POLITECNICO DI MILANO

Corso di Laurea Specialistica in Ingegneria Informatica

Dipartimento di Elettronica e Informazione

CARATTERIZZAZIONE DEGLI UTENTI TELEVISIVI PER LA SELEZIONE DI PUBBLICITÁ MIRATA

Relatore: Prof. Paolo Cremonesi
Correlatore: Ing. Roberto Turrin

Tesi di Laurea di:
Paolo Colombo, matricola 735088
Michele Giussani, matricola 749646

Anno Accademico 2010-2011

Sommario

In un settore dinamico e in rapido sviluppo come l'iPTV, la personalizzazione dei contenuti ed, in particolare, delle informazioni pubblicitarie, rappresenta uno strumento competitivo potenzialmente decisivo per i Content Provider.

Per proporre pubblicità mirate agli utenti, tuttavia, è necessario disporre delle informazioni demografiche ad essi associate, le quali, a causa del loro elevato valore e delle problematiche relative alla privacy, sono difficili da ottenere.

Lo scopo della tesi consiste nell'individuare nuove soluzioni per l'inferenza di tali informazioni, costruendo dei modelli per la caratterizzazione degli individui a partire dalle preferenze da essi espresse.

Per la valutazione dei risultati, oltre all'utilizzo di due strumenti tradizionali come l'Accuracy e la Recall, viene proposta una nuova metrica chiamata Lift, utile per il confronto delle performance ottenute applicando le metodologie di classificazione sperimentate rispetto all'assegnazione casuale degli utenti alle classi in base alla loro composizione.

Indice

Capitolo 1

Introduzione

La recente diffusione di tecnologie quali l'iPTV (IP Television) e di soluzioni software per la distribuzione di contenuti personalizzati (noti come sistemi di raccomandazione), offre interessanti opportunità in ambito commerciale. Una delle nuove funzionalità che sembra destare maggiore interesse è rappresentata dalla possibilità di mostrare agli utenti pubblicità mirata, il cosiddetto *Targeted Advertising* [30].

Al fine di presentare agli utenti messaggi pubblicitari mirati in base al loro profilo è necessaria una corretta segmentazione degli utenti in categorie [34]. Nell'ambito pubblicitario, tali categorie sono tipicamente definite sulla base di informazioni demografiche (in particolare sesso e fascia d'età); a seconda della categoria di appartenenza vengono quindi mostrate differenti pubblicità.

Non sempre però le informazioni demografiche degli utenti sono disponibili. Il nostro lavoro di tesi si sviluppa in questo scenario e ha come principale obiettivo quello di proporre e valutare soluzioni per la classificazione degli utenti in categorie di mercato - individuandone sesso ed età - a partire dal giudizio (rating) che gli stessi hanno espresso su un insieme di film visti.

1.1 Targeted Advertising

Il Targeted Advertising rappresenta un potenziale strumento di business per tutte quelle aziende che operano nel settore Internet e dei media. Il termine è accostato a quelle tecnologie e soluzioni informatiche utilizzate per ricavare ed analizzare le informazioni provenienti da un gran numero di utenti che usu-

fruiscono di particolari servizi, con l'obiettivo di proporgli pubblicità mirata. L'idea che sta alla base del Targeted Advertising è la volontà di proporre all'utente qualcosa che molto probabilmente coglierà la sua attenzione, studiando i suoi campi di interesse ed evitando così di subissarlo di messaggi promozionali che potrebbero essere ignorati in quanto ritenuti soltanto invasivi e fuori luogo [17].

La pubblicità mirata è per questo motivo una delle ragioni in grado di dar luogo ad una crisi dei media tradizionali e delle TV generaliste: infatti, il vero interesse della TV digitale non è tanto la possibilità di trasmettere un gran numero di canali diversi, né la capacità di offrire una definizione migliore delle trasmissioni, quanto piuttosto l'idea di inviare ad ogni spettatore il contenuto pubblicitario per lui più rilevante [4]. La stessa cosa accade, da molto prima dell'avvento della televisione digitale, nel mondo del web, dove un numero sempre crescente di aziende utilizza da tempo la pubblicità mirata: ogni volta che la pagina di un sito viene visualizzata, il servizio di Targeted Advertising ne identifica il contenuto ed il significato, fornendo automaticamente al visitatore un link pubblicitario coerente con il contesto delle pagine che sta visitando, oppure legato alle sezioni o categorie del sito [31]. Il grande vantaggio di questo tipo di pubblicità è la capacità di diventare visibile dove è più utile per i vari attori coinvolti: essa offre infatti una maggiore profondità informativa al contenuto che la ospita arricchendolo e spesso aggiungendo la componente commerciale come prezzi, offerte e sconti.

Nella realizzazione di soluzioni basate sul Targeted Advertising l'attenzione è posta in particolare sul concetto di *Lifestyle* [20], ovvero lo stile di vita di un individuo, inteso come insieme di attività, interessi, comportamenti, informazioni demografiche e quant'altro possa caratterizzarlo, che tiene conto sia delle diverse relazioni culturali, sociali, familiari, sia degli aspetti caratteriali di una persona. Al Lifestyle sono legati i gusti e i comportamenti abituali degli utenti, quindi anche i criteri con cui essi effettuano acquisti o usufruiscono di particolari servizi [21]. Proprio per questo motivo, esso viene considerato con particolare attenzione dai settori di marketing, che lo ritengono un valido strumento per impostare strategie vincenti dal punto di vista pubblicitario, visto che le fruizioni dei programmi televisivi degli utenti, abbinati alle loro informazioni anagrafiche, sono indicatori significativi dello stile di vita; la predizione del Lifestyle degli utenti rappresenta dunque uno dei punti chiave su cui è possibile basare l'impiego di pubblicità mirata.

1.1.1 Behavioural Targeting

Il Lifestyle può essere definito esplicitamente dall'utente, ad esempio inserendo le proprie informazioni demografiche nelle form di registrazione, oppure espresso implicito attraverso una serie di comportamenti ed attività svolte on-line. Nella seconda ipotesi si parla sempre più spesso di *Behavioural Targeting*.

Ricorrendo a questa tecnica, il profilo di un soggetto viene ricavato attraverso l'analisi approfondita delle attività e dei comportamenti da esso tenuti, in particolare per quanto riguarda le proprie abitudini di navigazione sul web. Per comportamento si intende l'insieme di azioni messe in atto da un certo utente durante la visita di un determinato sito, o la fruizione di un particolare servizio. Le pagine visitate, il tempo speso su ciascuna di esse, le keyword usate per raggiungerla, i click effettuati, sono tutti esempi che contraddistinguono i comportamenti degli utilizzatori e che vengono raccolte attraverso l'uso di dispositivi come cookies, web beacons, ad-server e analizzate attraverso opportune tecniche di data-mining [25].

Da tutte le informazioni raccolte si ricostruisce un profilo ampio e dettagliato degli utenti, cercando di risalire alle loro abitudini, interessi e scopi, in modo da poter rispondere meglio e più velocemente alle loro richieste, riuscendo anche ad anticiparne le necessità.

La pubblicità on-line rappresenta la fase conclusiva di questo monitoraggio e viene proposta agli utenti in base a quanto si riesce a ricostruire dal loro stile di navigazione ed utilizzo dei servizi, non riferendosi semplicemente al contesto, ovvero al sito all'interno del quale stanno navigando o al singolo programma che stanno guardando. Proprio in questo senso il Behavioural Targeting rappresenta uno strumento a supporto del Targeted Advertising [64]. Se una pubblicità contestuale classica, ad esempio, inserisce il banner di un hotel nel sito di un'agenzia di viaggi, le tecniche di Behavioural Targeting, riconoscendo che un utente è interessato all'argomento 'viaggi', permettono di inserire il banner di un albergo anche in un sito dedicato alla finanza o alla musica. I processi di individuzione del target improntati sulle tecnologie Behavioural, consentono infatti di seguire gli spostamenti di un utente, in modo da pubblicare gli annunci pubblicitari per i quali ha mostrato un evidente interesse anche in relazione a contenuti che non sono necessariamente correlati al prodotto o servizio che si vuole pubblicizzare.

Si raggiunge in questo modo la possibilità reale di veicolare messaggi sempre più personalizzati verso utenti che hanno effettivamente mostrato una certa propensione per un determinato tema. Le tecniche di Behavioural Targeting danno quindi luogo ad una pubblicità meno forzata nei confronti del consumatore, in quanto l'idea di fondo è che in realtà è l'utente stesso a richiederla implicitamente attraverso le attitudini che lo caratterizzano.

Il problema principale riguardo a questo tipo di soluzioni è costituito dalla intrusione nella privacy che viene compiuta per monitorare i comportamenti degli utenti, i quali solitamente non sono informati della presenza di meccanismi atti a tale scopo. Nonostante molti siti stiano compiendo passi in avanti da questo punto di vista, le modalità con cui si ricavano le informazioni e, soprattutto, per quali scopi esse saranno utilizzate, rimangono nella maggior parte dei casi poco chiare.

Negli ultimi anni, inoltre, si è assistito alla formazione di vere e proprie reti di siti, che condividono tra di loro le informazioni raccolte in modo da incrementare la quantità di dati disponibile per riconoscere i comportamenti dei propri visitatori. Spesso l'utente è ignaro dell'esistenza di tali 'cartelli' e si trova a fornire inconsapevolmente informazioni sulle proprie abitudini e preferenze di acquisto.

1.1.2 Lavori esistenti

Il dominio del Targeted Advertising si rivela essere particolarmente chiuso e riservato; ciò è testimoniato dal fatto che, nonostante l'alto interesse dimostrato dai Content Provider verso tale metodologia, il numero di pubblicazioni che descrivono soluzioni tecnologiche implementate in questo ambito sia limitato. Le soluzioni di maggiore interesse sono tipicamente basate su tecniche di data mining (es: SVM e alberi decisionali) e mostrate nel Capitolo 2, anche se la maggior parte dei lavori esistenti rimane piuttosto generica, senza riportare alcun dettaglio implementativo.

Il nostro lavoro si colloca in un percorso di ricerca già avviato da alcuni anni presso il Politecnico di Milano; in particolare, i primi ad essersi avvicinati all'ambito del Targeted Advertising sono Luigi Bianchi [7] e Fabio Granara [19]. Nel lavoro di Bianchi [7], vengono considerati tre tipi di Lifestyle degli utenti: il sesso, l'età e l'occupazione; per le predizioni degli utenti è stato utilizzato SVD, un algoritmo basato sulla fattorizzazione di matrici, facendo riferimento al

dataset fornito da Movielens. Lo studio viene condotto considerando soltanto tre generi di film, i preferiti, tra i diciotto per cui gli utenti hanno espresso una valutazione. Granara [19] ha invece approfondito il discorso su SVD considerando il Lifestyle dei generi ed utilizzando Movielens come dataset di riferimento. Nel suo lavoro viene inoltre preso in considerazione l'impiego delle Regole di Associazione, già ampiamente utilizzate nel mondo delle basi di dati e nel data-mining, come strumento di predizione dello stile di vita degli utenti.

In entrambi i casi la valutazione dei risultati è stata fatta basandosi su classiche metriche di accuratezza come Precision, Recall e curve di ROC, le quali hanno fornito risultati positivi che hanno dimostrato l'utilità delle tecniche impiegate, lasciando spazio a possibilità di approfondimento. In particolare Bianchi ha ottenuto valori di Precision e Recall prossimi al 75 % nei test sull'inferenza del sesso degli utenti [7], mentre le sperimentazioni compiute da Granara hanno portato a ricavare valori compresi tra il 54% e l'87% per la Precision e tra 24% 82% per la Recall [19].

1.2 Il contributo della tesi

L'obiettivo del nostro lavoro consiste nello sviluppo di algoritmi per la segmentazione di utenti in base all'età e al sesso, associando un Lifestyle a ciascuno di essi, in modo tale da fornire un supporto adeguato alle tecniche di Targeted Advertising. Come mostrato da Hawkins et al. in [21], infatti, le preferenze degli utenti sono legate alle loro caratteristiche demografiche quindi, una volta individuato il Lifestyle dell'utente, è possibile proporgli informazioni mirate e di conseguenza maggiormente efficaci.

Nell'ambito del nostro lavoro il punto di partenza per la determinazione dello stile di vita è rappresentato dalle preferenze espresse in relazione alla visione di una serie di film; in particolare, a partire da un insieme di rating forniti dall'utente, sono stati individuati i livelli di gradimento associati ai vari generi (es: commedia, horror, adult,ecc.). Questa scelta è giustificata dal fatto che il genere è stato dimostrato essere un buon discriminante per la caratterizzazione del sesso e dell'età dei soggetti [35, 36, 42]; l'importanza dell'analisi dei generi nella segmentazione del mercato è riconosciuta anche da compagnie specializzate, che li utilizzano per la caratterizzazione degli utenti in base al sesso e all'età [34].

Le informazioni relative agli utenti ed alle loro preferenze relative ai generi di film sono state ricavate da due basi di dati: oltre al dataset fornito da MovieLens, già utilizzato nei lavori precedenti, è stato infatti introdotto un nuovo dataset, elaborato da Yahoo!.

La classificazione degli utenti è stata effettuata ricorrendo alle tecniche basate sulle Regole di Associazione e SVD, precedentemente sperimentate da Bianchi [7] e Granara [19], andando ad approfondire le analisi realizzate in precedenza, in particolare effettuando test su un numero maggiore di parametri e sperimentando classi di età differenti. É stata inoltre utilizzata un'addizionale metodologia di apprendimento supervisionato, basata sulle Support Vector Machines. In aggiunta, ai singoli classificatori di base sono in seguito state applicate delle metodologie di combinazione (ad esempio tramite tecniche di Voting), nel tentativo di incrementare ulteriormente le performance delle soluzioni sviluppate.

Per valutare le metodologie proposte, inoltre, si è ritenuto utile definire una metrica aggiuntiva rispetto a quelle tradizionalmente utilizzate nei problemi di classificazione (Recall ed Accuracy), denominata Lift. Tale metrica consente di confrontare le prestazioni ottenute applicando una metodologia di Targeted Advertising con quelle ottenibili utilizzando una strategia di marketing tradizionale, in cui tutti gli utenti vengono assegnati alla classe maggiormente popolata. L'ottenimento di un valore di Lift superiore ad uno indica che le performance ottenute ricorrendo alla metodologia in esame nella fase di classificazione sono migliori rispetto al mancato utilizzo della stessa, ovvero ad un numero maggiore di utenti viene proposta pubblicità corrispondente al corretto Lifestyle.

Infine è stata introdotta una nuova tecnica, denominata Porting, che prevede la costruzione di un modello per la classificazione degli utenti a partire da informazioni demografiche contenute in un dataset e la sua applicazione a basi di dati che forniscono solo le preferenze degli utenti, senza però mettere a disposizione le informazioni demografiche necessarie per la costruzione di tale modello. L'applicazione di tale metodologia potrebbe portare benefici soprattutto a quei Content Provider che non abbiano a disposizione le informazioni demografiche relative ai propri utenti; tali informazioni, infatti, sono difficili da reperire e, perché il modello sia effettivamente efficace, devono essere in quantità significativa. Qualora un Content Provider volesse implementare soluzioni di Targeted Advertising potrebbe acquistare un modello

di classificazione sviluppato da altri soggetti, avendo così l'opportunità di sfruttare i benefici introdotti dal ricorso a tale strategia pubblicitaria, senza dover necessariamente effettuare una raccolta di informazioni demografiche degli utenti.

1.2.1 Aspetti innovativi

Alcuni degli argomenti da noi analizzati, realizzati e trattati in questo documento rappresentano novità significative nel campo del Targeted Advertising e lasciano intravedere anche interessanti prospettive per lo sviluppo della ricerca.

Uno degli aspetti maggiormente interessanti è la definizione della nuova metrica Lift che permette di valutare più approfonditamente i benefici ottenuti tramite il ricorso ad una metodologia di Targeted Advertising: ricorrendo ad essa è infatti possibile effettuare un'analisi più ampia delle performance ottenute, andando a cogliere aspetti che non potrebbero essere rilevati adoperando le tecniche di valutazione tradizionali.

Le Support Vector Machines inoltre, pur essendo state utilizzate in altri ambiti per risolvere problemi di varia complessità e natura, erano state scarsamente impiegate nel campo del Targeted Advertising. Tali tecniche hanno dimostrato un ottimo adattamento sia nel caso di problemi binari, sia nel caso di problemi multi-classe, con entrambi i dataset considerati.

Anche l'operazione di Porting da un dataset ad un altro non era mai stata valutata precedentemente con l'obiettivo di fornire un supporto nell'ottica della pubblicità mirata, né tantomento testata attraverso l'impiego di algoritmi come Regole di Associazione, SVD e SVM. Tale soluzione si presta ad avere sbocchi interessanti dal punto di vista commerciale, fornendo la possibilità di adottare modelli di dati particolari a chi ne è sprovvisto o non dispone di informazioni sufficienti per produrli, in modo da usufruire dei benefici che possono derivarne da un loro impiego.

1.2.2 Risultati rilevanti

Al termine del nostro studio possiamo evidenziare come le SVM rappresentino un potente strumento in grado di supportare la ripartizione degli utenti e la corretta assegnazione al loro profilo di appartenenza. Grazie alle SVM siamo riusciti infatti ad andare oltre risultati precedentemente raggiunti

nel campo del Targeted Advertising: in riferimento al dataset di Movielens e ai test sui sessi, siamo passati da Recall globali vicine al 70%, che venivano raggiunte nel corso di ricerche precedenti, a valori di tali metriche superiori al 79%. La nuova metrica Lift ha confermato ulteriormente l'efficacia delle SVM come strumento di supporto nel contesto sempre più articolato dei contenuti personalizzati e della pubblicità mirata, raggiungendo valori pari ad $1{,}0971$ per le predizioni sul sesso dell'utente, ovvero un miglioramento di circa il 10% rispetto all'uso di strategie pubblicitarie tradizionali. Per quanto riguarda i test sull'inferenza dell'età, invece i migliori risultati indicano un miglioramento di più del 13% rispetto ad una soluzione tradizionale.

1.3 Struttura della tesi

Nel Capitolo 2, dopo una breve panoramica su alcune soluzioni esistenti nell'ambito del Targeted Advertising, vengono illustrate le basi teoriche degli algoritmi da noi utilizzati, presentando in conclusione le principali metodologie per la combinazione dei classificatori. Il Capitolo 3 descrive l'organizzazione dei dati nei dataset Yahoo! e MovieLens, utilizzati come fonte di informazione per le nostre indagini. Nel Capitolo 4 si fa riferimento nel dettaglio alle metodologie di valutazione adoperate, con particolare attenzione posta sulla descrizione della nuova metrica da noi introdotta, oltre alla presentazione dello schema implementativo comune per le soluzioni basate sugli algoritmi utilizzati singolarmente. Le modifiche allo schema implementativo generale imposte dalle peculiarità tipiche di ciascun algoritmo sono invece descritte nel Capitolo 5. I risultati ottenuti nei test condotti per AR, SVD e SVM sono mostrati dettagliatamente nel Capitolo 6, all'interno del quale essi vengono anche confrontati tra di loro. Nel Capitolo 7 sono presentate due strategie da noi proposte per adattare le tecniche di combinazione di classificatori al nostro ambito operativo. Infine, il Capitolo 8 è dedicato alla tecnica del Porting, di cui sono descritti lo schema implementativo ed i risultati ottenuti tramite la sua applicazione.

Capitolo 2

Stato dell'arte

Indice

In questo capitolo sono presentate le basi teoriche relative agli algoritmi che saranno implementati all'interno del lavoro di tesi. Dopo una breve rassegna su alcune delle soluzioni sviluppate in letteratura nell'ambito del Targeted Advertising, nel Paragrafo 2.3 vengono descritti i tre algoritmi 'di base', che sono utilizzati separatamente per effettuare la caratterizzazione degli utenti. Il Paragrafo 2.4, invece, presenta alcune delle metodologie esistenti in letteratura per la combinazione dei classificatori, alcune delle quali saranno in seguito adoperate nel seguito per la costruzione di un unico classificatore composto, ottenuto dall'aggregazione dei tre algoritmi precedentemente utilizzati in modo separato.

2.1 Strategie di definizione dei Lifestyle

Nel corso degli anni sono state sviluppate sempre più strategie per la definizione del Lifestyle degli utenti allo scopo di individuarne le abitudini

di consumo e strategie di marketing efficaci. Nel seguito vengono presentate tre soluzioni, due sviluppate per il mercato statunitense ed una per il mercato italiano, tra le più diffuse e maggiormente utilizzate.

Claritas PRIZM NE [51] fornisce una classificazione dei proprietari di case statunitensi in 66 segmenti, ricavati analizzando sia le informazioni demografiche che le abitudini di consumo dei soggetti coinvolti nello studio. Nello sviluppo di tale analisi sono stati coinvolti più di 890 000 possessori di abitazioni; per individuare una mole così elevata di informazioni sono state effettuate apposite ricerche di mercato, combinate con le informazioni ricavate dal censimento della popolazione statunitense effettuato nel 2000 e dall'analisi di altre banche dati pubbliche come, ad esempio, i registri della motorizzazione civile.

Una delle caratteristiche che contribuiscono ad incrementare l'affidabilità del modello è che tutte le informazioni demografiche sono state fornite direttamente dai cittadini, i quali sono stati intervistati nell'ambito delle ricerche di mercato. I segmenti possono essere raggruppati utilizzando due metodologie: la prima consente di individuare i cosidetti Social Groups in base alla distribuzione sul territorio e al reddito degli utenti, la seconda, invece, prevede di considerare oltre alle informazioni basate sul reddito, anche l'età degli utenti e la presenza di bambini nel nucleo familiare.

VALS, acronimo di 'Values, Attitudes and Lifestyles', è un prodotto sviluppato, sin dagli anni '70, dalla SRIC-BI (Stanford Research Institute Consulting Business Intelligence) [60], che viene utilizzato per predire le abitudini di consumo degli utenti, in modo da consentire alle aziende la realizzazione di strategie di marketing particolarmente efficaci e maggiormente redditizie.

All'interno di VALS gli utenti vengono segmentati in otto profili, sulla base di due dimensioni: la dimensione verticale è influenzata dall'innovatività e delle risorse di cui dispongono gli utenti (reddito, educazione, doti di leadership, ecc.), mentre quella orizzontale è rappresentata dalle motivazioni che spingono il consumatore ad effettuare gli acquisti. Gli otto segmenti individuati sono i seguenti:

- *Innovators*: persone di successo, con alta autostima e molte risorse, economiche e non, a disposizione. Sono aperti alle nuove idee e tecnologie e hanno il ruolo di change leaders nella societa.

- *Thinkers*: motivati da ideali, sono persone mature e affidabili, conservative. Sono consumatori pratici, ricercano prodotti duraturi, funzionali e di valore.

- *Achievers*: motivati dal desiderio di realizzarsi, desiderano fare carriera e avere una famiglia. Sono conservativi, vivono vite tranquille e rispettano le autorità. Con molti desideri e necessita sono molto attivi nel mercato. Per loro l'immagine e importante e per questo desiderano prodotti di prestigio che li mettano in mostra.

- *Experiencers*: giovani, entusiasti e consumatori impulsivi. La loro energia cerca sfogo nello sport, nelle attività ricreative e sociali. Sono tipicamente avidi e spendono la maggior parte del loro reddito in articoli di moda e divertimento.

- *Believers*: motivati dagli ideali, sono persone conservative e convenzionali, il cui credo e fondato sulla religione, la famiglia, la comunità e la nazione. Sono consumatori prevedibili in quanto scelgono prodotti per la famiglia e acquistano brand classici e affermati.

- *Strivers*: il denaro per loro significa successo, ma non ne hanno abbastanza per raggiungerlo. Acquistano i prodotti che gli permettono di simulare uno status sociale superiore. Sono consumatori impulsivi, in dipendenza alla disponibilità economica del momento.

- *Makers*: persone pratiche, costruttive e auto sufficienti. Vivono in un contesto tradizionale, ad esempio la famiglia e il lavoro, e non mostrano interesse in quanto accade all'esterno. Non mostrano interesse nei beni materiali e nel lusso, per questo acquistano i prodotti essenziali.

- *Survivors*: hanno poche risorse a disposizione. Dovendo concentrarsi sul poter soddisfare le proprie necessita piuttosto che i desideri, non hanno un'evidente motivazione principale. Sono consumatori cauti e modesti. Apprezzano i beni di marca, specialmente se in saldo.

L'assegnazione degli utenti ad uno degli otto profili avviene sottoponendoli ad un questionario, i cui risultati sono in seguito analizzati per individuare quale Lifestyle sia maggiormente corrispondente alle reali abitudini di consumo dell'utente.

Per quanto riguarda il mercato italiano, invece, la maggiore strategia per la definizione dei Lifestyle è costitita da *Eurisko Sinottica* [53]. Si tratta di una indagine effettuata su campioni rappresentativi della popolazione italiana, svolta con l'obiettivo di apprendere informazioni sull'evoluzione socio-culturale, sui consumi e sull'esposizione ai mezzi di comunicazione dei cittadini con età maggiore di quattordici anni. Questa ricerca viene svolta a partire dal 1975, coinvolgendo per ogni anno 10 000 nuovi utenti, i quali vengono sottoposti ad interviste face-to-face presso le loro abitazioni. La rilevazione è destagionalizzata ed avviene in due periodi dell'anno, in modo da eseguire 5 000 interviste ogni semestre.

A partire dalle informazioni raccolte si applicano appositi strumenti di analisi, detti mappe di posizionamento, costituite da un certo numero di segmenti di mercato, ai quali sono associati differenti valori legati alle abitudini di consumo e culturali degli utenti. Attraverso l'uso di tali strumenti è possibile individuare quale sia il segmento che meglio si adatta alle abitudini di ciascun utente, derivandone di conseguenza il Lifestyle.

2.2 Alcune soluzioni di Targeted Advertising

In questo paragrafo sono mostrati esempi di metodologie operative sviluppate nell'ambito del Targeted Advertising. Nonostante il grande interesse emerso negli ultimi anni rispetto a tale disciplina, le soluzioni proposte in ambito accademico rappresentano un numero piuttosto limitato: il settore pubblicitario, infatti, costituisce un dominio relativamente 'chiuso', in cui la maggior parte delle ricerche viene sviluppata all'interno di progetti privati e con finalità commerciali.

Tutte le soluzioni presentate nel seguito comprendono una prima fase di caratterizzazione dell'utente, alla quale ne fa seguito una di individuazione delle informazioni pubblicitarie di maggiore rilevanza da proporre, le quali vengono selezionate a seconda della categoria a cui è stato assegnato l'utente.

Le informazioni necessarie per lo sviluppo di questo tipo di applicazioni riguardano principalmente le caratteristiche demografiche degli utenti e quelle

relative ai loro comportamenti ed alle preferenze da essi espressi e sono generalmente difficili da reperire. Tali informazioni, infatti, possono essere ricavate tramite ricerche di mercato, che però generalmente presentano costi elevati di realizzazione, oppure richiedendole direttamente agli utenti, ad esempio attraverso questionari o apposite procedure di registrazione on-line. Anche questa modalità di reperimento delle informazioni, però, non può essere considerata completamente affidabile, in quanto solitamente le persone vedono tali operazioni come delle minacce alla propria privacy e tendono a limitare al minimo la quantità di dati personali forniti. Inoltre, soprattutto per quanto riguarda le procedure di registrazione a servizi on-line, non vi è nessuna garanzia che le informazioni inserite siano effettivamente veritiere ed affidabili. L'elevato valore intrinseco di questo tipo di informazioni e la difficoltà nel loro reperimento fa si che il numero di dataset liberamente accessibili, e di conseguenza utilizzabili per l'analisi delle soluzioni implementative proposte in ambito accademico, sia limitato, rendendo più difficile la valutazione delle prestazioni realmente ottenute.

Un esempio di ricerca sviluppata in ambito aziendale è mostrato in [12], in cui viene presentato un brevetto sviluppato da Google che consente di costruire un modello basato sulle informazioni demografiche relative ai visitatori di youtube.com [61] e sul contenuto dei video per i quali essi mostrano maggiore interesse. Il primo passo della metodologia consiste nell'individuare un insieme di apprendimento formato da video che abbiano un elevato numero di visualizzazioni e nell'estrarre in modo automatico le informazioni associate al loro contenuto, tramite l'utilizzo di appositi riconoscitori. Successivamente si determinano quali sono gli utenti associati a tali video, ad esempio individuando coloro che lo hanno visualizzato ripetutamente o che hanno espresso un rating positivo, e se ne estraggono le informazioni anagrafiche, in particolare il sesso e la fascia di età a cui appartengono. Le informazioni demografiche così ricavate sono associate a quelle estratte automaticamente dal video ed entrambe vengono sottoposte ad un classificatore supervisionato, in questo caso SVM, il quale fornisce un modello per la predizione che consente di suggerire nuovi contenuti agli utenti in base al proprio profilo demografico. Trattandosi di una soluzione proposta in ambito privato e non di ricerca, non vengono forniti risultati derivati dall'applicazione di questo metodo.

Hu et al. [23] propongono un approccio per prevedere sesso ed età degli utenti a partire dai loro comportamenti durante la navigazione in Internet.

Sono presentate due differenti metodologie: nella prima, si individua, relativamente a ciascuna pagina appartenente all'insieme di apprendimento, la distribuzione di probabilità associata a ciascuna classe tra quelle previste, stabilendo quali siano le classi di utenti maggiormente interessate alla pagina in esame. Ricorrendo ad SVM si costruisce in seguito un modello basato sulle informazioni appena ricavate, per stabilire le distribuzioni di probabilità tra le varie classi per pagine non appartenenti all'insieme di apprendimento. Infine si utilizza tale modello per inferire le caratteristiche demografiche dell'utente a seconda delle pagine da esso visitate. Per rendere più efficace il metodo proposto si introduce una seconda metodologia per la categorizzazione, basata sull'individuazione di somiglianze tra gli utenti, partendo dall'assunzione che utenti con gusti simili visiteranno le stesse pagine e viceversa. Nella prima fase si individuano le somiglianze, applicando sulle informazioni relative alle navigazioni effettuate dagli utenti il metodo del *Latent Semantic Indexing*, basato su SVD. Successivamente si ricorre ad un'interpolazione lineare per la predizione degli attributi demografici a partire dalle somiglianze riscontrate tra le pagine visitate dagli utenti da classificare e quelle appartenenti all'insieme di apprendimento.

I risultati in termini di Recall e Precision per le due metodologie proposte sono riportati nella Tabella 2.1; le classi di età previste sono *Teenage*, comprendente gli utenti con meno di 18 anni, *Youngster* i cui membri hanno età compresa tra 18 e 24 anni, *Young*, i cui membri hanno età compresa tra 25 e 34 anni, *Mid-Age* i cui membri hanno età compresa tra 35 e 49 anni e *Elder*, di età superiore a 49 anni.

In Kim et al.[26] è presentato un sistema di raccomandazione in grado di proporre informazioni pubblicitarie personalizzate agli utenti di un negozio virtuale a seconda delle informazioni demografiche e degli acquisti effettuati in passato. La categorizzazione degli utenti viene eseguita applicando la tecnica del *tree induction*, che prevede la costruzione di alberi binari decisionali in cui ciascun nodo è costituito da un test a cui viene sottoposto l'utente; a seconda dei risultati ottenuti nei vari test, ciascun individuo è quindi assegnato ad una delle classi disponibili, che formano le foglie dell'albero. Una volta terminata la classificazione vengono prodotte delle regole che permettono di associare le categorie di avvisi pubblicitari e le classi di utenti in modo da individuare gli ad più rilevanti da proporre a ciascun utente.

La valutazione di questo metodo è stata effettuata raccomandando dei

CLASSE	Prima Metodologia		Seconda Metodologia	
	Recall	Precision	Recall	Precision
Maschi	0,711	0,707	0,810	0,791
Femmine	0,682	0,713	0,782	0,805
Teenage	0,323	0,361	0,457	0,471
Youngster	0,503	0,498	0,651	0,642
Young	0,492	0,486	0,642	0,632
Mid-Age	0,440	0,457	0,613	0,615
Elder	0,297	0,403	0,484	0,516

Tabella 2.1: *risultati ottenuti nell'inferenza del genere ed età da Hu et al in [23].*

brani musicali ad un campione di utenti, ai quali è stato chiesto di esprimere un giudizio da uno a cinque sul loro gradimento verso le raccomandazioni proposte. La media dei giudizi è pari a $3{,}34$, con una deviazione standard pari a $0{,}88$. Tali risultati sono migliori di quelli ricavati proponendo dei brani casuali agli utenti: applicando questa strategia si è infatti ottenuta una media delle valutazioni espresse pari a $2{,}90$, con deviazione standard di $0{,}76$.

In Yang et al.[65] è descritta una metodologia che prevede la suddivisione in gruppi degli utenti di un social network in modo da sottoporre informazioni pubblicitarie mirate a seconda del gruppo di appartenenza. I gruppi vengono costruiti cercando di individuare quelli utenti che hanno legami sufficientemente forti tra di loro che, di conseguenza, dovrebbero avere interessi e preferenze comuni. A partire da un database costruito analizzando sia le relazioni esplicitamente espresse dagli utenti (ad esempio i rapporti di amicizia), sia i legami impliciti, ricavati ad esempio dall'analisi dei log relativi all'utilizzo delle e-mail, si ottiene una rappresentazione grafica in cui i nodi rappresentano gli utenti, mentre i vertici indicano i legami tra di essi. I gruppi vengono quindi ricavati estraendo tutti gli utenti che hanno un certo numero di nodi 'vicini' in comune, individuati tramite l'algoritmo *k-nearest neighbor*; maggiore è il numero di nodi vicini in comune, maggiore sarà la coesione del gruppo e, di conseguenza, più elevata la probabilità che utenti appartenenti al gruppo abbiano i medesimi interessi. Nella fase di costruzione dei gruppi si ricorre ad un approccio bottom-up, simile all'algoritmo *apriori* utilizzato nell'ambito delle Regole di Associazione, in cui gli insiemi sono ricavati dall'unione di

sottoinsiemi di dimensione minore, fino ad arrivare alla costruzione di un insieme massimale, che non può quindi essere ulteriormente ampliato senza ridurre il livello di coesione tra gli utenti che lo compongono. L'associazione tra gruppi e spot pubblicitari da proporre viene invece effettuata analizzando il comportamento passato degli utenti: per ciascun gruppo si individuano le categorie per le quali essi hanno compiuto il maggior numero di transazioni, in modo da proporre informazioni pubblicitarie relative a prodotti appartenenti alle categorie più rilevanti per ognuno dei gruppi.

La valutazione è avvenuta costruendo una social network analizzando i log delle e-mail di un campione di studenti, oltre alle registrazioni relative ai prestiti di libri da essi effettuate presso la biblioteca dell'università in cui si è tenuto lo studio. L'ottanta per cento dei libri è stato quindi utilizzato come insieme di apprendimento, mentre il restante venti per cento ha svolto la funzione di insieme di test. La metrica di valutazione espressa è detta hit-rate e indica per ciascun libro la percentuale di utenti ai quali esso è stato raccomandato, tra coloro che effettivamente lo avevano preso in prestito. L'analisi dei risultati mostra che tale metrica raggiunge valori vicini all'ottanta per cento, nei test in cui si considerano tutte le registrazioni di prestiti effettuate nel periodo preso in considerazione.

Uno dei settori per i quali sono state implementate alcune soluzioni è quello relativo all'iPTV; ad esempio, in Spangler et al. [41] è presentato un sistema di Targeted Advertising basato sull'uso di più tecnologie di data mining combinate tra di loro, in cui l'obiettivo è effettuare la profilazione degli utenti a seconda di caratteristiche demografiche e comportamentali ricavate a partire dalla lista dei programmi televisivi da essi visionati. In particolare sono presi in considerazione la tipologia dei programmi, il numero di volte per cui ciascun contenuto è stato fruito e il tempo dedicato dall'utente alla sue visualizzazione. Le categorie di utenti sono costruite a partire dalle informazioni contenute nel database fornito da Nielsen Media Services [56]; in particolare vengono individuate cinque categorie di utenti, basate sulle loro caratteristiche demografiche, che sono mostrate in Tabella 2.2.
Per effettuare la classificazione si ricorre alle seguenti tecniche di data mining che operano separatamente:

- *k-nearest neighbor;*
- *Analisi discriminante;*

- *Regressione lineare;*
- *Alberi decisionali;*
- *Reti neurali;*
- *Classificatori bayesiani.*

Successivamente i risultati ottenuti dai singoli algoritmi sono combinati tra di loro tramite la tecnica del Voting (descritta nel Paragrafo 2.4.2), al fine di individuare la categoria di appartenenza di ciascun utente.

All'interno di questo studio viene proposta una metrica denominata Lift, calcolata come il rapporto tra la probabilità che un utente sia correttamente stimato come appartenente ad una classe e la probabilità che esso vi appartenga effettivamente. In questo modo è possibile valutare i miglioramenti introdotti per ciascuna classe rispetto all'uso di strategie che prevedono l'assegnazione casuale degli utenti. Un Lift maggiore di uno indica che il metodo applicato è effettivamente più performante rispetto all'assegnazione casuale. I risultati per le categorie di utenti previste sono riportati nella Tabella 2.2.

CLASSE	LIFT
Femmine di età compresa tra 18 e 34 anni	2,30
Femmine di età superiore a 55 anni	2,66
Maschi di età compresa tra 12 e 17 anni	2,73
Maschi di età compresa tra 18 e 34 anni	2,64
Maschi e Femmine di età compresa tra 2 e 11 anni	3,39

Tabella 2.2: *risultati ottenuti nell'inferenza del genere ed età da Spangler et al. in [41].*

Un'altra soluzione sviluppata nell'ambito della iPTV è il progetto iMEDIA presentato da Bozios et al. in [9], basato sull'uso di tecniche di *clustering* per l'individuazione di segmenti di utenti con gusti e preferenze simili, a cui proporre informazioni pubblicitarie mirate. Ogni utente dispone di un set-top box in grado di memorizzare le informazioni e comunicarle ai server centrali; a ciascun utente è inoltre richiesta la compilazione di un questionario contenente le proprie informazioni anagrafiche (sesso, età, luogo di residenza), le quali verranno in seguito combinate con i dati relativi ai contenuti fruiti, registrate dal set-top box, ed utilizzate per la suddivisione degli utenti in gruppi tramite algoritmi di clustering. I cluster così ricavati vengono analizzati tramite

apposite ricerche di mercato, con l'obiettivo di individuare le informazioni pubblicitarie più adatte per ciascuna tipologia di utenti. Le informazioni relative al cluster di appartenenza dell'utente sono memorizzate nel set-top box e comunicate al server durante la fruizione dei contenuti, in modo da poter proporre informazioni pubblicitarie mirate. Per questa soluzione in [9] non sono riportati risultati derivanti dall'effettiva applicazione su un campione di utenti, non permettendo quindi di valutarne appieno l'efficacia.

Il Lifestyle può essere espresso anche senza utilizzare informazioni anagrafiche ma, come proposto ad esempio in [48], cercando di individuare un insieme di categorie di maggior interesse per ciascun utente; in questa soluzione gli interessi degli individui vengono dedotti a seconda del contenuto delle foto da essi pubblicate su flick.com [59]. Le possibili categorie di interesse sono quelle proposte nell'ontologia ODP [52]; ciascun interesse è descritto tramite un vettore, detto Topic, ottenuto selezionando alcune pagine campione, dalle quali vengono estratti in modo automatico i termini più significativi. Successivamente per ogni immagine pubblicata vengono combinati i tag eventualmente generati dall'utente, con altri ottenuti in modo automatico usando il metodo Arista sviluppato da Wang et al.[49]. Per ogni immagine si valuta la somiglianza tra il vettore di tags che la descrivono e i vari Topic, utilizzando il metodo della *cosine similarity*; il risultato è anch'esso un vettore, detto Topic Distribution, formato da un numero di elementi pari al numero di categorie presenti nell'ontologia, i cui elementi rappresentano ciascuno il livello di somiglianza tra il contenuto dell'immagine e le varie categorie di interesse. Lo stesso procedimento è realizzato anche per i banner pubblicitari, in modo da costruire dei Topic Distribution ad essi associati. I Topic Distribution associati agli utenti e ai banner sono in seguito confrontati, in modo da proporre a ciascun utente solamente i banner che più si adattano alle categorie di interesse individuate.

Per valutare l'applicazione di questa metodologia sono state usate 5 000 fotografie, caricate da 25 utenti differenti ed associate ad un insieme di 20 milioni di prodotti provenienti dal catalogo di un sito di vendite on-line, suddivisi in 6 500 categorie. A ciascun utente sono quindi state proposte delle raccomandazioni relative a tali prodotti, chiedendo di esprimere un giudizio su di esse, scegliendo fra tre possibili risposte: 'ottimo', 'corretto' ed 'errato'.

La metrica di valutazione è stata espressa con la seguente formula:

$$\frac{(o + 0{,}5 * c)}{o + c + e}$$

in cui *o*, *c* ed *e* rappresentano il numero di giudizi ottimi, corretti ed errati. Tale metrica viene utilizzata per determinare i parametri che consentono le migliori performance per il metodo.

La soluzione presentata da Bae et al. [2], invece, si basa sull'uso delle cosiddette *Fuzzy Rules* per stabilire relazioni tra gli interessi degli utenti e le categorie di informazioni pubblicitarie da proporre. In una prima fase vengono individuate delle categorie di interesse e viene effettuata una analisi dei file di log associati a ciascun utente, in modo da stabilire il numero di accessi effettuati a siti relativi ad ognuna delle categorie predefinite, che viene memorizzato in un vettore. In seguito tali informazioni sono utilizzate per effettuare la segmentazione degli utenti, ricorrendo ad un particolare tipo di reti neurali, le *Self Organizing Map (SOM)*. Per ognuno dei segmenti individuati vengono costruite delle regole che, a partire dalle informazioni contenute nel vettore associato all'utente, creano delle strutture dette Fuzzy Set. Analoghe strutture vengono costruite per ognuna delle informazioni pubblicitarie, utilizzando apposite regole sviluppate da esperti di marketing. Nell'ultima fase un algoritmo calcola la somiglianza tra gli insiemi associati agli utenti e quelli relativi a ciascuna informazione pubblicitaria, in modo da individuare le più rilevanti da proporre.

Per valutare tale soluzione è stato calcolato il Click Through Rate (CTR) per alcune informazioni pubblicitarie selezionate, il quale è stato confrontato con il valore ottenuto proponendo i medesimi messaggi ad un gruppo di utenti casuale, senza cioè applicare il metodo di Targeted Advertising appena descritto. Le conclusioni sono positive, dato che si ottengono miglioramenti per il CTR relativamente a tutte le categorie di pubblicità proposte.

2.3 Algoritmi di base

L'obiettivo del nostro lavoro consiste nell'effettuare la caratterizzazione degli utenti ed individuarne di conseguenza il Lifestyle. Per raggiungere tale obiettivo abbiamo implementato tre soluzioni basate su altrettanti algoritmi:

le *Regole di Associazione*, la *Singular Value Decomposition* e le *Support Vector Machines*.

Come già accennato in precedenza i primi due algoritmi sono già stati utilizzati in questo ambito applicativo da Bianchi [7] e Granara [19]; il nostro studio prende spunto dalle soluzioni proposte per ampliare ulteriormente le analisi relative ai risultati ottenibili dall'applicazione di tali algoritmi. L'utilizzo dell'algoritmo basato sull'uso delle Support Vector Machine per la caratterizzazione degli utenti, invece, rappresenta una novità per questo ambito applicativo.

2.3.1 Le Regole di Associazione

Le *Regole di Associazione* (AR) sono una tra le metodologie di data mining più diffuse e approfonditamente studiate. Sono state introdotte all'inizio degli anni '90 nell'ambito della cosiddetta *market basket analysis*, una disciplina che si occupa dell'analisi delle abitudini di acquisto degli utenti, con l'obiettivo di fornire un supporto al management nella fase di definizione e valutazione delle strategie di marketing, pricing e product placing.

La nascita delle AR, unita allo sviluppo delle tecnologie per la gestione delle basi di dati e per la digitalizzazione delle informazioni, ha rappresentato un punto di svolta significativo per la market basket analysis, poiché ha permesso di sfruttare la grande quantità di informazioni raccolte nella fase di vendita, che fino a quel momento era pressoché inutilizzata a causa della mancanza di strumenti di analisi efficienti.

Nel corso degli anni questa tecnica ha trovato applicazione in numerosi ambiti, anche significativamente diversi da quelli previsti in origine come, ad esempio, la biologia [5], l'educazione [16], la sicurezza informatica [33]. Come vedremo in seguito, le AR possono essere applicate anche nell'ambito del Targeted Advertising, in particolare nella fase di caratterizzazione dell'utente e dell'inferenza del Lifestyle.

Le AR sono state presentate per la prima volta in un articolo di Agrawal [1], in cui vengono fornite una definizione formale ed un algoritmo per la ricerca delle regole in un database, oltre ad una serie di metriche volte alla valutazione della qualità delle regole individuate.
La definizione formale è la seguente:

dati un insieme *I*, contenente oggetti i detti *item*
$$I = \{ i_1, i_2, i_3, \ldots, i_m \}$$

ed un insieme di transazioni *D*, detto *database*
$$D = \{ t_1, t_2, t_3, \ldots, t_n \}$$

tali che ciascuna transazione $t_j \in D$ sia identificata in modo univoco
e contenga un insieme di item $i \in I$,

per regola di associazione si intende una implicazione nella forma

$$X \Rightarrow Y$$

$$\text{t.c.: } X \in I, \quad Y \in I, \quad X \cap Y = \emptyset.$$

L'insieme di item X è detto *antecedente* della regola, mentre l'insieme Y è detto *conseguente* della regola.

Le Regole di Associazione possono essere informalmente definite come un metodo per trovare relazioni di correlazione tra oggetti appartenenti a dataset di grandi dimensioni, ad esempio, basi di dati transazionali contenenti informazioni sulle vendite in un grande magazzino. Un'interpretazione intuitiva di questo tipo di regole è quella per cui se una transazione contiene un item A e vale la regola $A \Rightarrow B$, allora la transazione probabilmente conterrà anche l'item B.

Confidenza e supporto

Per effettuare la valutazione della qualità di una regola vengono proposte numerose metriche, tra le quali le due più significative sono:

- ***supporto***: indica la percentuale di transazioni appartenenti al database in cui è contenuta l'unione tra l'antecedente e il conseguente della regola. Nel caso della regola $A \Rightarrow C$, quindi, il supporto è espresso dal numero di transazioni che contengono *A* e *C* rispetto al totale delle transazioni esaminate.

$$\text{Supp}(A \Rightarrow C) = \frac{\text{numero di transazioni contenenti } A \text{ e } C}{\text{totale transazioni in } D} \tag{2.1}$$

In termini probabilistici, il supporto indica la probabilità "a priori" di individuare transazioni contenenti l'unione tra A e C nel database. Questa metrica fornisce quindi una misura della rilevanza statistica della regola.

- ***confidenza***: indica qual è la percentuale di transazioni che contengono l'antecedente, che contengono anche il conseguente. Nel caso della regola $A \Rightarrow C$, quindi, la confidenza può essere espressa come il rapporto tra il numero di transazioni che contengono A e C e il numero di transazioni che contengono A.

$$\text{Conf}(A \Rightarrow C) = \frac{\text{numero di transazioni contenenti } A \text{ e } C}{\text{numero di transazioni contenenti } A} \tag{2.2}$$

Dal punto di vista statistico, la confidenza esprime la probabilità condizionata di trovare C nelle transazioni che contengono A; questa metrica è quindi un indicatore della "forza" della regola.

Algoritmo per individuare le Regole di Associazione

L'algoritmo per la creazione delle AR si articola in due fasi: la prima detta di individuazione degli itemset frequenti, la seconda detta di generazione delle regole.

Nella prima fase l'algoritmo deve generare tutti i possibili insiemi di item all'interno della base di dati, calcolando per ognuno la frequenza con cui compare ed individuando i cosiddetti *itemset frequenti*, cioè gli insiemi aventi frequenza superiore rispetto ad un valore di soglia minimo prefissato.

Nella fase successiva, partendo da uno degli itemset frequenti individuati in precedenza, l'algoritmo deve creare tutte le possibili regole in cui compaiono item appartenenti all'insieme in esame, ed eliminare tutte quelle aventi confidenza e supporto minori rispetto ad un livello minimo; tale operazione deve esssere ripetuta per tutti gli itemset frequenti individuati nella prima fase, in modo da generare tutte le possibili Regole di Associazione per il database oggetto dell'analisi.

Esempio di applicazione.

Per comprendere meglio l'applicazione dell'algoritmo per la ricerca delle AR viene proposto un semplice esempio: a partire dal database transazionale

riportato nella Tabella 2.3(a) vengono in un primo momento estratti tutti gli itemset presenti, per poi selezionare solamente quelli frequenti, evidenziati in azzurro nella Tabella 2.3(b); come è possibile notare, sono considerati frequenti gli item con frequenza maggiore o uguale a 2.

Le AR individuate sono inserite nella Tabella 2.3(c), insieme ai relativi valori di supporto e confidenza, calcolati utilizzando le equazioni (2.1) e (2.2).

Transazione	Items
t_1	A, C,
t_2	A, B, C
t_3	B D
t_4	A, C, D
t_5	C, D

(a) *Un database transazionale*

Id	Itemset	Supporto
is_1	$\{A, B\}$	1
is_2	$\{A, C\}$	3
is_3	$\{A, D\}$	1
is_4	$\{B, D\}$	1
is_5	$\{C, D\}$	2
is_6	$\{A, B, C\}$	1
is_7	$\{A, C, D\}$	1

(b) *Itemset frequenti*

Regola	Confidenza	Supporto
$A \Rightarrow C$	1,00	0,60
$C \Rightarrow A$	0,75	0,60
$C \Rightarrow D$	0,50	0,40
$D \Rightarrow C$	0,66	0,40

(c) *Regole di Associazione.*

Tabella 2.3: *applicazione dell'algoritmo di ricerca delle AR.*

Ad esempio, per la regola $(A \Rightarrow C)$, i calcoli effettuati per stabilire i valori di confidenza e supporto sono i seguenti:

$$\text{Conf}(A \Rightarrow C) = \frac{\text{numero di transazioni contenenti } A \text{ e } C}{\text{numero di transazioni contenenti } A} = \frac{3}{3} = 1$$

$$\text{Supp}(A \Rightarrow C) = \frac{\text{numero di transazioni contenenti } A \text{ e } C}{\text{totale transazioni}} = \frac{3}{4} = 0{,}75$$

Inferenza del Lifestyle attraverso le Regole di Associazione

Come accennato in precedenza, l'applicazione del metodo basato sulle AR può rivelarsi uno strumento molto utile per inferire il Lifestyle di un utente. Nel nostro lavoro l'obiettivo consiste nell'individuare Regole di Associazione che consentano di categorizzare gli utenti a seconda dei loro generi di film preferiti: le regole che cercheremo di individuare saranno quindi del tipo:

$$\text{Movie\_Genre} \Rightarrow \text{Lifestyle}$$

in cui l'antecedente è costituito da uno o più generi di film preferiti dall'utente, mentre il conseguente indica il Lifestyle dell'utente in questione. Chiaramente il conseguente sarà formato da un unico elemento, dato che sarebbe inutile associare un utente a due o più Lifestyle contemporaneamente.
Vengono effettuati due tipi di analisi:

- Inferenza del *sesso* , in cui l'obiettivo è effettuare la suddivisione degli utenti nelle due categorie MASCHIO e FEMMINA.
- Inferenza dell'*età*, il cui obiettivo è categorizzare gli utenti a seconda della classe di età di appartenenza (ad esempio la classe GIOVANE comprende tutti gli utenti di età compresa tra 18 e 35 anni).

Esempi di regole utilizzate in queste analisi sono:

$$\text{Horror} \Rightarrow \text{MASCHIO}$$

$$\text{Horror, Crime} \Rightarrow \text{GIOVANE}$$

Non saranno invece considerate regole del tipo:

$$\text{Lifestyle} \Rightarrow \text{Movie\_Genre}$$

in quanto non risultano significative per le analisi effettuate nell'ambito di questo lavoro di tesi.

2.3.2 La Singular Value Decomposition

La *Singular Value Decomposition* (SVD) è un metodo per la fattorizzazione di una matrice di m righe e n colonne a coefficienti reali o complessi, che viene generalmente applicato nell'ambito di problemi quali la stima del rango

di una matrice, o la soluzione di sistemi lineari attraverso l'applicazione del metodo dei minimi quadrati [6]. Così come le AR, anche questa metodologia ha trovato applicazione nei più svariati ambiti, tra cui in particolare, gli algoritmi di raccomandazione.

La Singular Value Decomposition di una matrice M_{mxn} a coefficienti *reali* è la seguente:

$$M = U\ \Sigma\ V^T \tag{2.3}$$

Applicando questa fattorizzazione è possibile quindi esprimere la matrice originaria *M* come il prodotto di tre matrici:

- *U* è una matrice *ortogonale (mxm)*, cioè una matrice quadrata tale che:

$$U^T\ U = I_m \qquad e \qquad U\ U^T = I_m$$

- V^T è la *trasposta* di una matrice *ortogonale (nxn)* tale che:

$$V^T\ V = I_n \qquad e \qquad V\ V^T = I_n$$

- Σ è una matrice *(m x n)* così costruita:

$$\Sigma \quad = \begin{bmatrix} \sigma_1 & 0 & 0 & 0 \\ 0 & \sigma_2 & 0 & 0 \\ 0 & 0 & \cdots & 0 \\ 0 & 0 & 0 & \sigma_n \end{bmatrix} = \quad diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$$

Gli elementi σ_i, posizionati sulla diagonale di Σ, sono tutti maggiori o uguali a 0 e sono detti *valori singolari*; il numero di valori singolari diversi da zero è pari al *rango* della matrice *M*.

Le colonne di *U* sono gli *autovettori* della matrice $M \cdot M^T$, e sono dette *vettori singolari sinistri*, mentre le colonne di *V* sono gli *autovettori* della matrice $M^T \cdot M$ e sono denominate *vettori singolari destri*. I vettori singolari (sinistri e destri) e quindi le matrici *U* e *V* non sono univocamente determinati, mentre i valori singolari e la matrice Σ sono univocamente determinati.

Valori e vettori singolari

In una matrice rettangolare *M* i vettori ed i valori singolari possono essere paragonati agli autovettori e agli autovalori di una matrice singolare: è infatti possibile utilizzarli per "ricostruire" la matrice *M* di partenza, attraverso la seguente formula:

$$M = \sum_{j=1}^{n} \sigma_j \, \vec{u_j} \, \vec{v_j}^T \tag{2.4}$$

in cui $\vec{u_j}$ e $\vec{v_j}^T$ sono dei vettori rappresentanti rispettivamente la *j*-esima colonna della matrice *U* e la *j*-esima riga della matrice V^T dell'equazione (2.3).

Inoltre, se la matrice *M* ha rango *r*, è possibile ordinare i suoi valori singolari in ordine decrescente in modo tale che le prime *r* posizioni sulla diagonale di Σ siano occupate dagli elementi diversi da 0.
La matrice $\overline{\Sigma}$ che si ottiene applicando l'ordinamento dei valori singolari è quindi:

$$\overline{\Sigma} = \; diag \; (\underbrace{\sigma_1, \sigma_2, \ldots, \sigma_{r-1}, \sigma_r}_{>0}, \underbrace{\sigma_{r+1} \ldots \sigma_n}_{=0})$$

$$\text{t.c. } \sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_{r-1} \geq \sigma_r > 0$$

E' possibile osservare come il contributo dei valori singolari uguali a 0 nella "ricostruzione" della matrice *M* sia nullo: l'equazione (2.4) è quindi equivalente alla seguente:

$$M = \sum_{j=1}^{r} \sigma_j \, \vec{u_j} \, \vec{v_j}^T \tag{2.5}$$

che si ottiene non considerando i valori singolari uguali a 0 che, dopo l'ordinamento, occupano le ultime *n-r* posizioni sulla diagonale di $\overline{\Sigma}$.
Per lo stesso motivo l'equazione (2.3) può essere espressa come:

$$M = U_r \Sigma_r {V_r}^T \tag{2.6}$$

in cui:

- Σ_r è una matrice diagonale di dimensione *r* ottenuta eliminando dalla matrice $\overline{\Sigma}$ gli *n-r* valori singolari uguali a 0.

- U_r è una matrice di dimensione *mxr*, ottenuta eliminando le ultime *n-r* colonne dalla matrice *U* ottenuta dall'equazione (2.3).
- $V_r{}^T$ è una matrice di dimensione *rxn*, ottenuta eliminando le ultime *n-r* righe dalla matrice V^T dall'equazione (2.3).

Approssimazione lower-rank di una matrice con SVD

É possibile dimostrare [15] che, considerando una matrice rettangolare a coefficienti reali *M* di rango *r* e dato un intero *k*<*r*, la migliore approssimazione di *M* di rango *k* è la matrice M_k:

$$M_k = U_k \Sigma_k V_k{}^T \tag{2.7}$$

ottenuta calcolando le matrici $\overline{\Sigma}$, *U* e V_T con il metodo della Singular Value Decomposition e considerando solo i primi *k* valori singolari (in ordine decrescente) di Σ, e le prime *k* colonne di *U* e *V*.

In accordo con le equazioni (2.4) e (2.5), la matrice M_k può essere espressa anche nella forma:

$$M_k = \sum_{j=1}^{k} \sigma_j \ \vec{u_j} \ \vec{v_j}^T \tag{2.8}$$

in cui σ_j, $\vec{u_j}$ e $\vec{v_j}^T$ indicano rispettivamente un valore singolare, un vettore singolare sinistro ed un vettore singolare destro di *M*.

Uso di SVD per l'inferenza del Lifestyle

La Singular Value Decomposition può essere applicata anche nell'ambito degli algoritmi di raccomandazione e, dopo aver effettuato alcune modifiche alle strutture dei dati utilizzate, anche per l'inferenza del Lifestyle.

Il punto di partenza per l'analisi è una matrice URM o UGM, la quale viene scomposta nelle matrici *U*, Σ, *V* (equazione 2.3). In questo modo le colonne della matrice *U* costituiscono gli autovettori associati agli utenti, mentre le colonne della matrice *V* rappresentano gli autovettori associati agli items da raccomandare. Le matrici prodotte vengono poi approssimate utilizzando un valore *k* (equazione 2.7): se tale valore è minore del rango della matrice di partenza, il modello così generato è in grado di ridurre il rumore e generare raccomandazioni.

É possibile dimostrare che la lista delle raccomandazioni per un utente può essere ottenuta semplicemente moltiplicando il vettore associato all'utente per V_k e ${V_k}^T$ [40], come nella seguente equazione:

$$\text{reccList}_i = \text{userGenres}\ V_k\ {V_k}^T$$

Nel nostro lavoro l'obiettivo non consiste nell'effettuare raccomandazioni, bensì nel classificare gli utenti a seconda del Lifestyle di appartenenza: è quindi necessario utilizzare un dataset appositamente creato, ottenuto concatenando una matrice UGM, rappresentante i generi preferiti dagli utenti, con una matrice che ne descriva il Lifestyle. A partire dal dataset così creato si genera il modello e si ottiene il Lifestyle di ciascun utente con la seguente equazione:

$$\text{Lifestyle}_i = \text{userGenres}\ V_{cl}\ {V_{cl}}^T$$

in cui *cl* è il numero di classi in cui si vogliono suddividere gli utenti, mentre V_{cl} è la matrice ottenuta mantenendo solo le ultime *cl* colonne della matrice modello *V*. Grazie a questo accorgimento si vanno a considerare soltanto quelle colonne che nel dataset originario sono associate alla descrizione del Lifestyle,

A seconda del valore $Lifestyle_i$ ottenuto ciascun utente viene assegnato ad una classe di appartenenza. Nel nostro lavoro l'assegnazione sarà effettuata applicando la cosiddetta *binarizzazione*, in cui si fa riferimento ad un valore di soglia *threshold* prefissato e gli utenti sono assegnati a classi differenti a seconda se il loro valore $Lifestyle_i$ superi o meno quello di *threshold*.

2.3.3 Support Vector Machines

L'algoritmo SVM, acronimo di Support Vector Machines, identifica una serie di metodologie di apprendimento supervisionato utilizzabili come valido elemento di supporto in relazione a problematiche di classificazione e regressione di particolari pattern. Le Support Vector Machines sono state sviluppate negli AT&T Bell Laboratories da Vladimir Vapnik e dai suoi colleghi [8, 13, 37, 38], la cui ricerca, dopo un primo approccio di stampo prettamente teorico, ha avuto nel tempo una decisa inclinazione applicativa.

L'algoritmo Support Vector alla base delle SVM è in realtà una generalizzazione non lineare dell'algoritmo Generalized Portrait sviluppato in precedenza dallo stesso Vapnik [46, 45] e come quest'ultimo risulta inquadrabile nella

statistical learning theory o teoria di Vapnik-Chervonenkis [47, 44]. La teoria VC definisce le proprietà degli algoritmi di apprendimento che devono essere in grado di 'imparare' sulla base delle informazioni fornite da un set di dati di 'esempio' e consentire una generalizzazione in relazione a dati 'nuovi', per mezzo di quanto viene messo in atto durante la prima fase. Le SVM operano esattamente in questa maniera.

Definizione

Una SVM è definita come un classificatore binario in grado di apprendere il confine fra esempi appartenenti a due diverse classi, proiettandoli in uno spazio multidimensionale e cercando in esso un iperpiano di separazione. L'obiettivo è quello di scegliere l'iperpiano in grado di massimizzare la sua distanza dagli esempi di training più vicini procedendo attraverso una serie di step successivi. Le SVM presentano inoltre alcune proprietà generali, in particolar modo:

- improbabile overfitting (ovvero adattamento a caratteristiche esclusive del campione di training)
- capacità di saper gestire dati con molte caratteristiche descrittive
- compattamento dell'informazione contenuta nel dataset in input

Fase di apprendimento (learning)

In caso di classificazione di dati appartenenti a due sole classi (problema binario), il processo di apprendimento si sviluppa seguendo una logica ben precisa.

L'approccio alla SVM avviene innanzitutto con il passaggio di una serie di dati di ingresso (detti esempi di training) alla macchina (o classificatore), in modo da ricavare un preciso modello di classificazione. Gli esempi di training sono dati pre-classificati, ovvero dotati in partenza di una label che identifica la loro appartenenza ad una determinata classe. Nei problemi binari, gli esempi di training sono generalmente etichettati con la label '+1' in caso di appartenenza alla classe positiva (o prima classe) e con la label '-1', qualora fossero legati alla classe negativa (o seconda classe).

Il training set viene dunque definito come l'insieme dei punti:

$$(x_1, y_1), ..., (x_l, y_l), x_i \in R_n, y_i \in \{-1, +1\}$$

presi da una distribuzione sconosciuta $P(x, y)$.

Dato un insieme di funzioni dette 'funzioni di soglia' definite come:

$$\{f_\alpha(x) : \alpha \in \Lambda\}, f_\alpha : R_n \longmapsto \{-1, +1\}$$

dove Λ è un insieme di parametri reali, si vuole trovare una funzione f_α in grado di minimizzare l'errore teorico

$$R(\alpha) = \int |f_\alpha(x) - y| P(x, y) dx dy$$

Le funzioni f_α sono chiamate ipotesi, l'insieme $\{f_\alpha : \alpha \in \Lambda\}$ viene chiamato spazio delle ipotesi e si indica con H; l'errore teorico rappresenta quanto è buona un'ipotesi nel predire la classe y_i di un punto x_i.

La distribuzione di probabilità $P(x, y)$ non è nota, quindi non è possibile calcolare l'errore teorico $R(\alpha)$. Come è possibile intuire, è però disponibile un campione di $P(x, y)$: il tranining set. Si può quindi calcolare un'approssimazione di $R(\alpha)$, ovvero l'errore empirico $R_{emp}(\alpha)$:

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} |f_\alpha(x_i) - y(i)|$$

La *Legge dei grandi numeri* garantisce che l'errore empirico converge in probabilità all'errore teorico, per cui si cerca di minimizzare l'errore empirico piuttosto che quello teorico.

La dimensione VC dello spazio di ipotesi H (detta anche dimensione del classificatore f_α) è un numero naturale che corrisponde al più grande numero di punti che possono essere esaminati e separati in tutti i modi possibili dall'insieme di funzioni f_α; la dimensione VC può essere finita, ma anche infinita.

Dato quindi un insieme di punti l, se per ognuna delle 2^l possibili classificazioni $(-1, +1)$ esiste una funzione f_α che assegna correttamente le classi, allora si dice che l'insieme di punti viene separato dall'insieme di funzioni.

La teoria della convergenza uniforme in probabilità, sviluppata da Vapnik

e Chervonenkis, fornisce anche un limite alla deviazione dell'errore empirico dall'errore teorico. Fissato quindi un η con $0 \leq \eta \leq 1$ vale la disuguaglianza:

$$R(\alpha) \leq R_{emp}(\alpha) + \sqrt{\frac{(h(log(2l/h) + 1) - log(\eta/4))}{l}}$$

dove h è la dimensione VC di f_α ed l il numero di datapoints che saranno passati in ingresso come input.

Per ottenere l'errore teorico minimo, bisogna minimizzare sia l'errore empirico sia il rapporto tra la dimensione VC e il numero di punti (h/l). L'errore empirico è solitamente una funzione decrescente di h, quindi, per un dato numero di punti, esiste un valore ottimale della dimensione VC (che risulta dipendente da un trade-off tra R_{emp} e h/l). L'algoritmo SVM è in grado di risolvere il problema minimizzando sia la dimensione VC sia il numero di errori sul training set.

L'obiettivo è quindi quello di scegliere la learning machine con l'$R(\alpha)$ minimo (processo questo detto di Structural Risk Minimization o SRM), andando a cogliere quel subset di funzioni (a cui sono legati i parametri α e h), attraverso il training dei classificatori che consentono di ottenere un valore minimo (o comunque molto basso) di tale parametro.

Applicabilità dell'algoritmo SVM

Esistono diverse modalità di applicazione dell'algoritmo SVM a seconda che si faccia riferimento a *macchine lineari* che operano su *dati linearmente separabili* o su *dati non linearmente separabili* e a *macchine non lineari*. Nel primo caso si considera un sottoinsieme lineare delle sopracitate funzioni di tipo f_α, mentre per le macchine del secondo tipo si fa riferimento ad un sottoinsieme di tipo non lineare.

Macchine lineari e dati linearmente separabili

Il caso più semplice di applicabilità dell'algoritmo SVM è quello di macchine lineari che operano su dati separabili, alla base delle quali vi sono le linee guida dettate dal seguente:

Teorema 1 *Se si considera un insieme di punti in R^n, scelto un punto come origine, i restanti punti possono essere divisi da un set di iperpiani se e solo se i vettori posizione*

dei restanti punti sono linearmente indipendenti. Da questo teorema si evince che la dimensione VC di un set di iperpiani relativo ad un insieme di punti in R^n è $n + 1$.

Si suppone in questo caso di avere degli iperpiani che separano gli esempi negativi (che supporremo identificati con label '-1') da quelli positivi (che supporremo identificati da label '+1'). I punti che giacciono sull'iperpiano sono quelli che soddisfano l'equazione:

$$\mathbf{w} \cdot \mathbf{x} + b = 0$$

dove $\mathbf{w}$ è un vettore normale all'iperpiano, $||\mathbf{w}||$ è la norma euclidea, $\frac{b}{||\mathbf{w}||}$ la distanza perpendicolare dell'iperpiano dall'origine. Come accennato precedentemente, si cerca di trovare l'iperpiano il cui margine dagli esempi ad esso più vicini è massimo e che è riconosciuto come iperpiano migliore.

Per calcolare il margine dell'iperpiano, ovvero l'incremento dello 'spessore', nel caso di dati linearmente separabili, si cerca di determinare la coppia $(\mathbf{w}, b)$ dell'iperpiano (che identifica i parametri della funzione da scegliere), considerando i seguenti vincoli:

$$\mathbf{w} \cdot \mathbf{x}_i + b \geq +1 \; con \; \mathbf{x}_i \in classe \; 1 \; (y_i = +1) \tag{2.9}$$

$$\mathbf{w} \cdot \mathbf{x}_i + b \leq -1 \; con \; \mathbf{x}_i \in classe \; 2 \; (y_i = -1) \tag{2.10}$$

che vengono combinati nel set di disuguaglianze:

$$y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0 \; \forall i$$

In questo caso, lo spazio delle ipotesi è formato dall'insieme di funzioni:

$$h : f_{\mathbf{w},b} = sign(\mathbf{w} \cdot \mathbf{x} + b)$$

I punti che soddisfano la (2.9) stanno sull'iperpiano

$$H_1 : \mathbf{x}_i \cdot \mathbf{w} + b = 1$$

detto *plus panel*, mentre quelli che soddisfano la (2.10) stanno sull'iperpiano

$$H_2 : \mathbf{x}_i \cdot \mathbf{w} + b = -1$$

detto *minus panel*. H_1 e H_2 sono paralleli tra loro (hanno la stessa normale) e nessun esempio di training cade nello spazio compreso tra essi.

Il margine è definito come $|d_+ + d_-|$, dove d_+ è l'esempio positivo appartenente al *plus panel* più vicino all'iperpiano di separazione, mentre d_- è l'esempio negativo appartenente al *minus panel* più prossimo all'iperpiano di separazione.

I punti per cui vale la (2.9) giacciono sull'iperpiano H_1 secondo la normale $\mathbf{w}$ e con distanza perpendicolare all'origine pari a $\frac{|1-b|}{||\mathbf{w}||}$. Allo stesso modo, i punti per cui vale la (2.10) giacciono sull'iperpiano H_2 secondo la normale $\mathbf{w}$ e con distanza perpendicolare all'origine pari a $\frac{|-1-b|}{||\mathbf{w}||}$. Si deduce quindi che $d_+ = d_- = \frac{1}{||\mathbf{w}||}$ e che il margine $|d_+ + d_-|$ è pari a $\frac{2}{||\mathbf{w}||}$.

In caso di dati linearmente separabili, lo scopo dell'SVM è quello di trovare tra tutti gli iperpiani che classificano correttamente il training set quello che ha norma minima $||\mathbf{w}||^2$, cioè margine massimo rispetto ai punti del training set. Mantenere questa norma piccola consente anche di basarsi su una dimensione VC piccola.

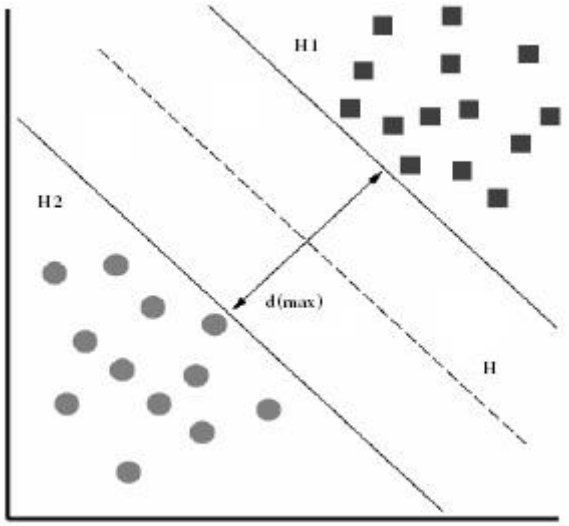


Figura 2.1: *Funzionamento di SVM: H è l'iperpiano di separazione, d è il margine pari a $|d_+ + d_-|$, con d_+ e d_- che giacciono rispettivamente sugli iperpiani H_1 e H_2*

Per costruire l'iperpiano ottimo, bisogna quindi classificare correttamente i punti del training set nelle due classi $y_i \in \{-1, 1\}$, usando la più piccola norma di coefficiente $\mathbf{w}$.

Possiamo tradurre il tutto in un problema di ottimizzazione lineare; il nostro obiettivo è infatti quello di minimizzare

$$\Phi(\mathbf{w}) = \frac{1}{2}||\mathbf{w}||^2$$

con $\mathbf{w}, b$ soggetti al vincolo

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \ i = 1, ..., l \tag{2.11}$$

Si passa quindi ad una diversa formulazione del problema, basandosi sulla tecnica dei moltiplicatori di Lagrange, sia perché in questo modo si possono esprimere i vincoli in maniera più agevole da gestire, sia perché i dati di training appariranno solo sotto forma di prodotto scalare tra vettori.

Si ottiene dunque:

$$L_p(\mathbf{w}, b, \Lambda) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{i=1}^{l} \alpha_i y_i(\mathbf{x}_i \cdot \mathbf{w} + b) + \sum_{i=1}^{l} \alpha_i \tag{2.12}$$

dove α_i rappresenta un moltiplicatore lagrangiano (punto stazionario), uno per ogni vincolo della forma $y_i(\mathbf{x}_i \cdot \mathbf{w} + b) - 1 \geq 0$, mentre $\Lambda = (\alpha_1, ...\alpha_l)$ è il vettore dei moltiplicatori di Lagrange non negativi relativi ai vincoli imposti dalla (2.11). Si cerca quindi di minimizzare L_p con riferimento a $\mathbf{w}$ e a b, massimizzandola contemporaneamente rispetto a $\Lambda \geq 0$.

Differenziando quindi la (2.12) ed impostando il risultato uguale a zero si ottiene:

$$\frac{\partial L(\mathbf{w}, b, \Lambda)}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^{l} \alpha_i y_i \mathbf{x}_i = 0 \tag{2.13}$$

$$\frac{\partial L(\mathbf{w}, b, \Lambda)}{\partial b} = \sum_{i=1}^{l} \alpha_i y_i = 0$$

Dalla (2.13) si ottiene:

$$\mathbf{w}^* = \sum_{i=1}^{l} = \alpha_i^* y_i \mathbf{x}_i \tag{2.14}$$

L'iperpiano ottimo può essere scritto come una combinazione lineare dei vettori del training set:

$$\mathbf{w}^* = \sum_i \alpha_i^* y_i \mathbf{x}_i = \Lambda^* y\, \mathbf{x}$$

da cui si ottiene

$$f(x) = \mathbf{w}^* \cdot \mathbf{x} + b = \Lambda^* y\, \mathbf{x} \cdot \mathbf{x} + b$$

Essendo questo un problema quadratico complesso, è possibile arrivare allo stesso risultato, ma in maniera più agevole, risolvendo il problema duale dato dalla massimizzazione di:

$$L_d = \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j\, \mathbf{x}_i \cdot \mathbf{x}_j \tag{2.15}$$

soggetta ai vincoli

$$\alpha_i \geq 0 \quad i = 1, ..., l \tag{2.16}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0 \tag{2.17}$$

Riformulando il problema in questa maniera, i vincoli imposti dalla (2.11) sono sostituiti da vincoli sui moltiplicatori e i vettori del training set appaiono solamente nella forma di prodotti interni tra vettori.

Risolvendo la (2.15) con i vincoli (2.16) e (2.17) si determinano i moltiplicatori Λ. L'iperpiano ottimo dipende dalla (2.14), mentre il classificatore è dato da:

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i^* (\mathbf{x} \cdot \mathbf{x}_i) + b^*)$$

per ogni vettore $\mathbf{x}_i$.

Tutti i punti $\mathbf{x}_i$ della soluzione per cui il corrispondente moltiplicatore α_i è strettamente maggiore di zero, vengono detti *support vector* e si trovano su uno dei due sopracitati iperpiani H_1 e H_2. I rimanenti punti del training set hanno il corrispondente α_i uguale a zero e non influenzano il classificatore.

I *support vector* sono i punti critici del training set e sono i più vicini all'iperpiano di separazione; se invece tutti gli altri punti venissero rimossi o

spostati senza oltrepassare H_1 e H_2 e l'algoritmo di apprendimento venisse ripetuto, si otterrebbe esattamente lo stesso risultato.

Macchine lineari e dati non linearmente separabili

In un contesto di dati linearmente non separabili esistono punti in posizione anomala rispetto agli altri punti della stessa classe, i quali possono sforare nel dominio opposto. In questo caso, si considera una costante di scarto ξ tanto maggiore quanto più lontani sono i punti anomali.

L'iperpiano separatore per un insieme di punti non linearmente separabili ha distanza $\frac{-b}{||\mathbf{w}||}$ dall'origine e viene determinato dai support vector. Il punto in posizione anomala dista $\frac{-\xi}{||\mathbf{w}||}$ dalla classe di appartenenza.

Nel caso di dati non linearmente separabili si procede quindi modificando la (2.9) e la (2.10) aggiungendo delle variabili ξ indicative del margine di errore:

$$\mathbf{x}_i \cdot \mathbf{w} + b \geq 1 - \xi_i \quad per \quad y_i = +1$$

$$\mathbf{x}_i \cdot \mathbf{w} + b \leq -1 + \xi_i \quad per \quad y_i = -1 \tag{2.18}$$

che, combinate, danno luogo a

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad i = 1, ..., l$$

In questo modo il vincolo ammette una certa tolleranza ξ_i agli errori. Affinché un punto del training set venga mal classificato, il corrispondente ξ_i deve superare l'unità, mentre la sommatoria dei vari ξ_i costituisce un upper bound relativamente agli errori di training.

L'obiettivo diventa ora quello di minimizzare $\frac{1}{2}||w||^2 + C(\sum_{i=1}^{l} \xi_i)^k$, dove C è un parametro assegnato dall'utente e rappresenta la penalità di errore; più alto è C, maggiore è la penalità di errore. Questo è un altro problema di ottimizzazione, per cui dobbiamo minimizzare

$$\Phi(\mathbf{w}, \Xi) = \frac{1}{2}||w||^2 + C(\sum_{i=1}^{l} \xi_i)^k \tag{2.19}$$

con $\mathbf{w}$, b e Ξ vincolati da:

$$y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - \xi_1, \quad \xi_i \geq 0 \quad i = 1, ..., l \tag{2.20}$$

L'algoritmo SVM cerca quindi di minimizzare $||w||$ e allo stesso tempo separare i punti dati, commettendo il minimo numero di errori possibile.

La soluzione al problema di ottimizzazione dato dalla (2.19) e soggetto al vincolo (2.20) si trova nello stesso modo del caso linearmente separabile, giungendo alla formulazione lagrangiana

$$L(\mathbf{w}, b, \Lambda, \Xi, \Gamma) = \frac{1}{2}||w||^2 - \sum_{i=1}^{l} \alpha_i[y_i(\mathbf{w} \cdot \mathbf{x}_i + b) - 1 + \xi_i] - \sum_{i=1}^{l} \gamma_i \xi_i + C(\sum_{i=1}^{l} \xi_i)^k \tag{2.21}$$

in cui i moltiplicatori $\Lambda(\alpha_1, ..., \alpha_l)$, $\Gamma(\gamma_1, ..., \gamma_l)$ sono associati alla (2.20). La (2.21) deve essere minimizzata rispetto a $\mathbf{w}$, b, Ξ e massimizzata rispetto a $\Lambda \geq 0$ e $\Gamma \geq 0$.

Supponendo $k = 1$ come nel caso di dati separabili, per semplificare i calcoli, si arriva ad una riformulazione duale del problema simile alla (2.15) che implica la massimizzazione di

$$L_d = \sum_{i=1}^{l} \alpha_i - \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j \mathbf{x}_i \mathbf{x}_j$$

facendo riferimento ai vincoli

$$0 \leq \alpha_i \leq C, \;\; i = 1, ..., l$$

La soluzione è identica a quella mostrata per i dati linearmente separabili, con la differenza che, in questo caso, è necessario considerare il vincolo sui moltiplicatori che sono limitati superiormente da C. Si avrà quindi un classificatore del tipo:

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i^* (\mathbf{x} \cdot \mathbf{x}_i) + b^*)$$

Macchine non lineari

Nel caso di dati non linearmente separabili si procede in un'altra maniera. L'obiettivo è quello di 'aumentare' il numero di dimensioni dei datapoint, in modo tale da potersi basare su di un nuovo criterio per poter effettuare una corretta separazione.

Si suppone in primo luogo di mappare i dati in un nuovo spazio euclideo H (che può anche avere un numero infinito di dimensioni, ma comunque presenta almeno una dimensione superiore a quella dei dati), utilizzando la funzione:

$$\Phi : R^d \longmapsto H$$

in modo tale che nell'algoritmo di training si vengano a trovare prodotti scalari in H nella forma $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$. Uno spazio di dimensione maggiore causa però seri problemi di calcolo, perché l'algoritmo di apprendimento deve lavorare con vettori di grandi dimensioni.

Per ovviare a questo problema e per evitare di esplicitare che cosa sia Φ (è difficile farlo soprattutto se H ha un numero infinito di dimensioni), si introduce una funzione *Kernel* che restituisce il prodotto delle immagini dei suoi due argomenti

$$K(\mathbf{x}_i \cdot \mathbf{x}_j) = \Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$$

Grazie a tale funzione, si è in grado di evitare di eseguire il prodotto esplicito tra le immagini dei vettori. E' dunque possibile sostituire K all'interno dell'algoritmo ed ignorare la forma esplicita di Φ.

Se si procede sostituendo nell'algoritmo $x_i \cdot x_j$ con $K(x_i, x_j)$, si dà luogo ad una Support Vector Machine che lavora in H e produce un risultato nella stessa quantità di tempo che impiegherebbe se lavorasse con i dati originali non mappati. Quello che si cerca di fare è sostanzialmente costruire un classificatore lineare in uno spazio differente: si mappa la variabile di ingresso in uno spazio di dimensione maggiore e si lavora attraverso una classificazione lineare in questo nuovo scenario.

Un punto x viene mappato in un vettore (detto vettore di feature) tramite la funzione Φ:

$$\mathbf{x} \rightarrow \Phi(\mathbf{x}) = (a_1\Phi_1(\mathbf{x}), a_2\Phi_2, ...) \tag{2.22}$$

dove gli a_i sono numeri reali e le Φ_i sono funzioni reali.

Si applica quindi lo stesso algoritmo del caso non separabile, sostituendo la variabile x con un nuovo vettore di feature $\Phi(x)$. Tenendo conto della (2.22), la funzione di decisione diventa quindi

$$f(x) = sign(\Phi(\mathbf{x}) \cdot \mathbf{w}^* + b^*) = sign(\sum_{i=1}^{l} y_i \alpha_i^* \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_j) + b^*)$$

Sostituendo i prodotti scalari $\Phi(\mathbf{x}_i) \cdot \Phi(\mathbf{x}_j)$ con una funzione kernel

$$K(\mathbf{x}, \mathbf{y}) \equiv \Phi(\mathbf{x}) \cdot \Phi(\mathbf{y}) = \sum_{i=1}^{\infty} a_n^2 \Phi_n(\mathbf{x}) \cdot \Phi_n(\mathbf{y})$$

si ottiene

$$f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i^* K(\mathbf{x}, \mathbf{x}_i) + b^*)$$

Non sempre esiste una coppia (H, Φ); per capire quando tale coppia esiste è necessario basarsi sulle cosiddette

Condizioni di Mercer: *Esiste un mapping definito da Φ in uno spazio H e una conseguente funzione kernel K se e solo se esiste una funzione $g(x)$ tale che*

$$\int g(x)^2 dx$$

sia finita e che

$$\int K(\mathbf{x}, \mathbf{y}) g(\mathbf{x}) g(\mathbf{y}) dx dy \geq 0$$

Quindi in sostanza una funzione kernel che soddisfa le condizioni di Mercer rappresenta un prodotto scalare in uno spazio delle feature generato da una qualche trasformazione non lineare.
Esempi di funzioni kernel sono:

- Lineare
$$K(\mathbf{x}, \mathbf{y}) = \mathbf{x} \cdot \mathbf{y} \tag{2.23}$$

- Polinomiale
$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y} + 1)^p$$

- Radial Basis Function

$$K(\mathbf{x}, \mathbf{y}) = e^{(-\gamma||\mathbf{x}-\mathbf{y}||^2)} \tag{2.24}$$

Il kernel (2.23) viene preferito nei casi di input set di dimensioni di gran lunga superiori a quelle delle categorie a cui i vari esempi fanno riferimento, mentre, negli altri casi in cui non sussiste questa peculiarità, solitamente si utilizza il kernel (2.24), il quale fornisce invece una funzione di classificazione radiale gaussiana.

SVM e problemi multi-classe

Fino ad ora abbiamo parlato di SVM supponendo di operare in un contesto binario, dove i dati devono essere classificati come appartenenti ad una delle due classi sotto esame (la classe positiva indicata con label '$+1$' e la classe negativa indicata con label '-1'). E' possibile però utilizzare l'algoritmo in altri contesti e adattarlo esplicitamente a situazioni e scenari più complessi, dove il numero delle classi in base alle quali suddividere i dati può essere superiore a due. Una situazione di questo tipo (*problema multi-classe*) si affronta generalmente secondo due modalità: la *Pairwise* e la *One against all* classification.

SVM e Pairwise Classification (Round Robin Classification)

La tecnica di Pairwise classification consiste nello scompattare il problema multi-classe originale in una serie di sottoproblemi binari a ciascuno dei quali viene assegnato un classificatore che agisce considerando solo gli esempi di training della coppia di classi presa in considerazione, ignorando tutte le altre. In questo modo si fa derivare la predizione del problema multi-classe dalle predizioni dei set di classe binari.

La tecnica viene preferita ad altre che si prefiggono lo stesso obiettivo, in quanto la fase di learning di un certo numero di sottoproblemi di un singolo round robin viene affrontata molto più efficientemente rispetto al learning messo in pratica con altre tecniche che cercano di arrivare al medesimo scopo.

La binarizzazione di tipo round robin infatti trasforma un problema multiclasse c in $\frac{c(c-1)}{2}$ problemi binari, dove $< i, j >$ saranno le classi prese in

considerazione per ogni set di classi $\{w, z\}$, con i variabile da 1 a $c-1$ e j variabile da $i+1$ a c.

Il classificatore per $< i, j >$ viene 'istruito' tramite gli esempi di classe i e di classe j, ignorando tutti gli esempi delle altre classi. L'algoritmo di learning utilizzato prende il nome di *base learner*.

Dopo aver suddiviso il problema, è necessario capire come ricondurre ad una valutazione finale le informazioni raccolte per ogni esempio di test. Generalmente tale valutazione viene fatta nel seguente modo: ciascuno dei classificatori base learner che opera sui sottoproblemi binari determina a quale delle due classi è più probabile che il nuovo esempio appartenga, attribuendogli, in base a questa decisione, un determinato punteggio. Quando ciascun classificatore ha terminato la valutazione, si considerano i vari punteggi e si assegna all'esempio la label della classe che ha accumulato il maggior numero di punti. In caso di parità invece, si assegna la label della classe i cui esempi sono presenti in percentuale maggiore nel set di training.

Bisogna tener conto del fatto che, essendo ogni classificatore forzato ad esprimere per ogni esempio una valutazione e quindi un punteggio, è probabile che in alcuni casi vengano assegnate delle label in maniera errata. Questa situazione può verificarsi generalmente quando esiste una certa somiglianza tra classi. Infatti, nel caso comune in cui le classi sono molto diversificate tra loro (potremmo dire quindi, nel caso di classi indipendenti), questo problema non si verifica, in quanto i classificatori che hanno 'imparato' da esempi della classe sono certamente in grado di assegnare al nuovo esempio un punteggio superiore a quello fornito dagli altri base learner che danno una valutazione errata (e, come specificato sopra, forzata).

Fino ad ora si è supposto che il procedimento di distinzione della classe i da j porti allo stesso risultato del caso in cui si prenda in esame il processo inverso da j ad i.

Quando il base learner è simmetrico non ci sono problemi; alcuni algoritmi di learning però operano in maniera diversa: essi scelgono una classe come elemento di default e semplicemente 'imparano' le regole per la distinzione dalla classe opposta. E' chiaro che in uno scenario simile la situazione cambia e si ottengono risultati diversi nei due casi (i, j) e (j, i).

La questione si può risolvere attraverso l'esecuzione di un doppio round robin (double round robin), in cui classificatori separati sono istruiti sia per problemi di tipo (i, j) che per problemi di tipo (j, i). Con questa tecnica, un

problema multiclasse c viene scomposto in $c(c-1)$ two-class problem, uno per ogni paio di classi (i, j), dove i varia da 1 a c, mentre j deve essere diverso da i. Gli esempi di classe i sono utilizzati come esempi positivi, quelli di classe j come esempi negativi.

E' definita come *class penalty* la relazione:

$$\pi_{b|f} = \frac{g_{b|f}(c, n)}{f(n)}$$

dove $f(n)$ (esprimibile come λn^p, con p in questo caso uguale ad uno) indica il tempo necessario affinché un learning classifier apprenda da un training set di n esempi, mentre $g_{b|f}(c, n)$ rappresenta la complessità totale relativa all'utilizzo di un learner con complessità temporale $f(n)$ su di un problema con un numero di classi pari a c, che è stato binarizzato utilizzando l'algoritmo b.

In sostanza, la class penalty è fondamentalmente la valutazione delle performance di un algoritmo su di un multiclass c-problem binarizzato, rispetto alla sua performance su un singolo two-class problem con uno stesso numero di esempi n nel training set.

Per un singolo round robin che fa riferimento ad un algoritmo di learning lineare, la class penalty vale:

$$\pi_{r|f_1}(c) = c - 1$$

dove $f_p(n) = \lambda n^p$ con $\lambda \geq 0$ e $p = 1$.

Per un singolo round robin che fa riverimento ad un algoritmo di learning super-lineare (ovvero con $f_p(n) = \lambda n^p$ con $p > 1$), la class penalty sarà data da:

$$\pi_{r|f_p}(c, n) = \begin{cases} c - 1 & \textit{se c è pari} \\ c & \textit{se c è dispari} \end{cases}$$

One against all/Unordered classification

Una strategia alternativa di binarizzazione porta all'adozione della tecnica di *One against all/Unordered classification*.

Le premesse sono le stesse del caso di pairwise classification: si vuole risolvere un problema multiclasse c, andando a dividerlo in un determinato numero di two-class problems, per poi rielaborare le informazioni ricavate

dai vari sottoproblemi in un'unica decisione finale. Per la precisione, la one against all classification trasforma un problema multiclasse c in un numero c di sottoproblemi binari (i, j). Questi sono costruiti considerando come positivi gli esempi di classe i e, come negativi, tutti gli esempi delle rimanenti classi j, con i che varia da 1 a c e con j diverso da i. Il concetto chiave è quindi quello di distinguere ogni volta una classe da tutte le altre.

Come accade nell'altra modalità, per ogni nuovo esempio verrà assegnato dai classificatori un determinato punteggio. Al termine, l'esempio verrà etichettato con l'identificativo della classe che ha ottenuto il punteggio maggiore.

Anche per l'Unordered classification viene definita una class penalty pari a:

$$\pi_u(c) = c$$

2.4 Metodologie per la combinazione di classificatori

Nel seguente paragrafo vengono presentate le principali tecniche per la combinazione di classificatori ottenuti dell'applicazione di differenti metodologie di data mining, il cui obiettivo consiste nell'ottenere un miglioramento delle performance rispetto all'utilizzo separato dei singoli metodi.
Gli obiettivi da perseguire attraverso l'uso della combinazione di classificatori sono in particolare:

- migliorare le performance complessive della fase di classificazione;
- sfruttare le peculiarità di ciascun classificatore individuandone i punti di forza, ovvero le situazioni in cui il classificatore mostra una maggiore abilità rispetto agli altri metodi utilizzati;
- ridurre il tasso di errore complessivo, confrontando le stime fornite da ciascun classificatore e correggendo quelle considerate errate.

Come evidenziato da Ho in [22], utilizzando tali tecniche si passa da un problema in cui l'obiettivo è individuare il classificatore che garantisca le migliori performance, ad un nuovo problema finalizzato alla costruzione di un insieme di classificatori e all'individuazione della tecnica più efficiente per combinarli.

Le principali metodologie per combinare tra di loro le stime fornite dai singoli classificatori di base sono le seguenti:

- algoritmi per la *selezione* di classificatori,
- algoritmi per la *fusione* di classificatori
- algoritmi basati sull'uso di *meta-classificatori*.

Tutte le metodologie proposte prevedono una fase iniziale di scelta dei classificatori da combinare, i quali dovranno essere in seguito aggregati secondo modalità specifiche per ciascuna di esse.

La tecnica della *selezione dei classificatori* prevede che l'assegnazione degli item alla classe di appartenenza sia realizzata mediante l'utilizzo di un unico classificatore, scelto tra tutti quelli a disposizione. É quindi necessaria una fase preliminare in cui definire i parametri da tenere in considerazione per determinare quale classificatore utilizzare.

Applicando la *fusione di classificatori*, invece, si combinano tra di loro i risultati di tutti i classificatori di base a disposizione, ad esempio effettuando la media delle varie stime fornite. A differenza di quanto avviene nella selezione vengono prese in considerazione tutte le predizioni fornite dai classificatori di base, i quali sono quindi tutti coinvolti nella fase di caratterizzazione degli item.

Sono state anche implementate soluzioni intermedie, come ad esempio quella proposta da Kuncheva in [28], in cui l'insieme degli item viene inizialmente suddiviso in regioni e, per ciascuna di esse, ad ogni classificatore viene assegnato un peso proporzionale alla propria abilità nel classificare gli item appartenenti alla regione considerata. Le varie stime fornite sono quindi combinate tra di loro, come previsto dalla metodologia della fusione, tenendo conto dei pesi assegnati ai vari classificatori nella fase precedente, in modo da ottenere la classe a cui assegnare l'item. Anche in questo caso tutti i classificatori considerati partecipano alla fase di caratterizzazione, con un'incidenza sulla stima finale diversa a seconda della regione a cui appartiene l'item da classificare.

Gli algoritmi che fanno uso dei cosiddetti *meta-classificatori* sono caratterizzati dall'introduzione di un livello intermedio nel processo di classificazione. Le stime di base sono infatti restituite ad un secondo livello di classificatori, i quali le utilizzano come meta-dati per effettuare la caratterizzazione finale

degli item. I classificatori di secondo livello, quindi, operano su insiemi di apprendimento formati sia dagli attributi associati agli item, sia dalle stime fornite dai classificatori di base. Il procedimento può essere ripetuto per un numero arbitrario di livelli, anche se risultati soddisfacenti possono essere ottenuti anche limitandosi ad operare su due livelli. L'obiettivo di questo tipo di algoritmi è quello di risolvere gli errori commessi nel corso delle prime fasi, cercando allo stesso tempo di eliminare eventuali incongruenze tra le stime fornite dai classificatori di base.

2.4.1 La selezione di classificatori

Questi algoritmi sono stati presentati per la prima volta da Dasarathy e Sheela in [14]. La selezione dei classificatori può essere realizzata *staticamente*, suddividendo il dominio di applicazione in sotto-regioni prima della fase di classificazione degli item, oppure *dinamicamente*, selezionando di volta in volta il classificatore più adatto a seconda dei risultati delle stime appena effettuate.

Selezione statica

In questa famiglia di algoritmi, prima dell'inizio della fase di classificazione si creano delle regioni di competenza e si elegge un classificatore responsabile dell'assegnazione di una classe agli item che vengono identificati come appartenenti a tale regione. L'algoritmo deve quindi essere in grado di stabilire quale dei classificatori sia il più adatto per ognuna delle regioni di competenza definite, oltre ad individuare correttamente a quale regione appartenga ciascun item. La suddivisione dell'insieme degli item di partenza in regioni può avvenire basandosi su dati storici ricavati da passate classificazioni, oppure affidandosi a tecniche di clustering.

Il *clustering* è una tecnica di data-mining che prevede di analizzare gli attributi di ciascun item del dominio di partenza e suddividere quest'ultimo in un numero k di partizioni, dette cluster, basandosi su relazioni di somiglianza tra gli attributi. Ciascuna delle partizioni individuate rappresenta, quindi, una regione di competenza a cui è necessario assegnare un classificatore.

La fase di selezione prevede che gli item appartenenti alla regione siano stimati utilizzando tutti i classificatori di base a disposizione, in modo da calcolare per ciascuno di essi l'accuratezza delle stime fornite; il classificatore che fornisce le migliori performance nella predizione della classe di appartenenza

degli item della regione viene designato come responsabile. Nella fase di classificazione vera e propria, l'item in esame viene assegnato ad una delle regioni di competenza, e successivamente stimato dal classificatore responsabile, designato in precedenza.

Questa metodologia, descritta da Kuncheva in [27], prende il nome di *Clustering and Selection*; sono state sviluppate anche metodologie più complesse in cui si utilizzano algoritmi di clustering maggiormente sofisticati e criteri più restrittivi per la selezione del classificatore responsabile, come proposto da Liu e Yuan in [32].

Selezione dinamica

Utilizzando la selezione dinamica, invece, la scelta del classificatore avviene durante la fase di classificazione, selezionando di volta in volta il miglior algoritmo.

Inizialmente viene costruito un insieme di item considerati 'vicini' a quello in esame, ad esempio utilizzando l'algoritmo *k-nearest neighbors*. Si seleziona quindi il classificatore di base che fornisce le migliori prestazioni nella stima degli item appartenenti all'insieme dei vicini, che verrà utilizzato per assegnare una classe all'item in esame. La selezione del classificatore può avvenire basandosi sulla percentuale di item classificati correttamente, come proposto da Woods et al. in [63], oppure calcolando la probabilità di successo di ciascun classificatore, pesata sul numero di stime effettuate; questo metodo è stato utilizzato da Giacinto e Roli in [18].

Le tecniche appena descritte prevedono di selezionare il classificatore e successivamente classificare l'item. Una alternativa è quella di stimare l'item con tutti i classificatori di base e solo in seguito effettuare la selezione del classificatore. L'item i viene inizialmente stimato con un classificatore in modo da ottenere la stima $c_{i,n}$; in seguito si individuano i k item più vicini ad i che sono stati stimati come appartenenti alla classe $c_{i,n}$ e si calcola l'accuratezza del classificatore, intesa come la percentuale di classificazioni corrette. Il classificatore che consente di ottenere le migliori prestazioni viene quindi utilizzato per la stima di i. Anche per questa tipologia di algoritmi sono state descritte metriche più complesse per la selezione del classificatore, le quali considerano anche la distanza dei k punti 'vicini' rispetto all'item i; un esempio di tale tecnica è stato proposto da Giacinto e Roli in [18].

2.4.2 La fusione di classificatori

Questa metodologia prevede di tenere in considerazione tutte le stime fornite dai classificatori di base; nel seguito vengono illustrate le caratteristiche principali dei più utilizzati algoritmi per la fusione di classificatori.

Voting

Gli algoritmi basati sul *Voting* sono stati i primi ad essere utilizzati per integrare tra loro risultati provenienti da differenti classificatori. Fare riferimento al lavoro di Kuncheva [29] per una analisi approfondita delle varie tipologie di Voting.

Questi algoritmi prevedono che ognuno dei classificatori di base a disposizione effettui la stima della classe di appartenenza dell'item i. Al termine della prima fase di classificazione, per ciascuna classe *c* e per ciascun item *i*, si calcola un punteggio pari al numero di metodi base che hanno stimato *i* come appartenente a *c*. L'item viene quindi assegnato alla classe che ha ottenuto il punteggio più alto: maggiore sarà il punteggio ottenuto, più elevata sarà l'attendibilità della classificazione.

Avendo a disposizione un numero *N* di classificatori e definita $\hat{S}_n$ la predizione fornita dal classificatore *n-esimo*, il punteggio ottenuto da ciascuna classe *c* nella classificazione di un item si calcola come:

$$\text{Score}_c = \sum_{n=1}^{N} s_{c,n}$$

in cui $s_{c,n}$ è così definito:

$$s_{c,n} = \begin{cases} 1 & \text{se } \hat{S}_n = c \\ 0 & \text{altrimenti} \end{cases} \tag{2.25}$$

La classe $\hat{c}$ da assegnare all'item si individua determinando il massimo punteggio ottenuto:

$$\text{Class} = \hat{c}$$
$$t.c : \text{Score}_{\hat{c}} = max(\text{Score}_c)$$

Il punteggio minimo da raggiungere per assegnare l'utente ad una determinata classe varia a seconda del livello di attendibilità richiesto: generalmente si richiede un numero di voti superiore alla metà del numero di classificatori di base utilizzati. Qualora nessuna classe dovesse raggiungere il punteggio minimo prefissato l'item non verrebbe dunque assegnato.

Una variante del voto a maggioranza, detta *Unanimity voting*, si applica quando il livello di attendibilità richiesto sia particolarmente elevato; questo metodo impone che l'assegnamento avvenga solo se una classe raggiunga un punteggio pari a N, cioè se tutti i classificatori considerati siano concordi nella stima. Nel caso di classificazione binaria, cioè quando il numero di possibili classi a cui assegnare un item è pari a 2, si utilizzano generalmente un numero dispari di classificatori, per evitare che si verifichino situazioni di pareggio tra le due classi, e si richiede che una delle due classi raggiunga la maggioranza dei voti a disposizione.

Utilizzando gli algoritmi appena descritti, il voto di ciascun classificatore di base contribuisce alla determinazione del punteggio finale ottenuto dalle classi nella stessa misura, senza considerare le peculiarità di ciascun classificatore come, ad esempio, l'accuratezza complessiva o l'abilità nell'individuare utenti appartenenti ad una specifica classe. Sono state quindi proposte delle varianti alla metodologia di base, in cui a ciascun classificatore viene assegnato un peso, determinato a seconda delle caratteristiche e dalle performance da esso ottenute. In queste metodologie, dette *Weighted voting*, il punteggio ottenuto da una classe c nella classificazione di un item i viene ottenuto moltiplicando ciascun voto per il peso assegnatoli e sommando i voti pesati dei singoli classificatori.

$$\text{Score}_c = \sum_{n=1}^{N} s_{c,n} * \omega_n$$

in cui ω_n è il peso assegnato al classificatore n, mentre $s_{c,n}$ è calcolato come nell equazione (2.25).

Il peso da assegnare a ciascun classificatore può essere determinato secondo vari criteri, ad esempio impostando pesi proporzionali all'accuratezza ed alla probabilità di successo del classificatore, o inversamente proporzionali alla probabilità di errore.

Naive Bayes Combination

Questo metodo utilizza il concetto di probabilità *a posteriori*, considerando tutte le possibili combinazioni di stime fornite dai classificatori di base, e cercando di individuare la classe da assegnare ad un item a seconda della combinazione di stime ottenuta.

In un prima fase, tutti i classificatori di base vengono utilizzati per stimare gli item contenuti in un insieme di apprendimento, dei quali siano note le classi di appartenenza. Per ogni classe *c* e per ogni possibile combinazione $\hat{\Sigma} = (\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N)$ delle stime fornite dai classificatori di base, si calcola la probabilità a posteriori che un item appartenga a *c*, se i classificatori di base hanno fornito una combinazione di stime $\hat{\Sigma}$.
Per il Teorema di Bayes, tale probabilità vale:

$$P\big(c|\hat{\Sigma}\big) = P\big(c|\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N)\big) = \frac{P\big(c\big)P\big((\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N|c\big)}{P\big(\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N\big)}$$

in cui $P\big(c\big)$ è la probabilità che un item preso a caso dal campione di apprendimento appartenga alla classe *c*, calcolata come il rapporto tra il numero degli item appartenenti alla classe *c* e il numero totale di item presenti nel campione. Inoltre, se i classificatori sono tutti indipendenti tra di loro vale che:

$$P\big((\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N|c)\big) = \prod_{n=1}^{N} P\big(\hat{S}_n|c\big)$$

Il valore di $P\big(\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N\big)$ non dipende dalla classe *c*, quindi:

$$P\big(c|(\hat{S}_1, \hat{S}_2, \ldots, \hat{S}_N)\big) = P(c) \prod_{n=1}^{N} P\big(\hat{S}_n|c\big) \tag{2.26}$$

Nella fase di classificazione degli item si ricava la combinazione di stime $\hat{\Sigma}$, utilizzando i classificatori di base e si assegna la classe $\hat{c}$ per la quale si ottiene il massimo valore $P\big(\hat{c}|\hat{\Sigma}\big)$, calcolato attraverso l'equazione (2.26).

Behavior Knowledge Space

Questo metodo, presentato da Huang in [24], è molto simile alla Naive Bayes Combination descritta in precedenza. Anche in questo caso la prima fase consiste nella stima degli item contenuti nell'insieme di apprendimento

con i classificatori di base. Successivamente, si costruisce una lock-up table formata da un numero di righe pari a quello delle classi presenti ed un numero di colonne pari a quello delle possibili combinazioni $\hat{\Sigma}$ delle stime fornite dai classificatori di base. Per ciascuna classe *c* e per ciascuna combinazione $\hat{\Sigma}$ si calcola il numero di item dell'insieme di apprendimento appartenenti a *c*, per i quali i classificatori di base hanno restituito la combinazione di stime $\hat{\Sigma}$ e lo si inserisce all'interno della lock-up table nella cella corrispondente.

La Tabella 2.4 è un esempio di lock-up table ricavata per la stima di item utilizzando due possibili classi c_1 e c_2, ed i classificatori S_1 e S_2.

	$\hat{S}_1 = c_1$ $\hat{S}_2 = c_1$	$\hat{S}_1 = c_1$ $\hat{S}_2 = c_2$	$\hat{S}_1 = c_2$ $\hat{S}_2 = c_1$	$\hat{S}_1 = c_2$ $\hat{S}_2 = c_2$
c_1	127	58	92	14
c_2	35	62	131	194

Tabella 2.4: *esempio di look-up table per l'applicazione del metodo BKS.*

Nella fase di caratterizzazione degli item di cui non è nota la classe di appartenenza, si ricava la combinazione delle stime, applicando tutti i classificatori di base. In seguito si individua la colonna della lock-up table associata alla combinazione ottenuta per l'item, il quale viene assegnato alla classe corrispondente alla riga con il maggior numero di utenti.

	$\hat{S}_1 = c_1$ $\hat{S}_2 = c_1$	$\hat{S}_1 = c_1$ $\hat{S}_2 = c_2$	$\hat{S}_1 = c_2$ $\hat{S}_2 = c_1$	$\hat{S}_1 = c_2$ $\hat{S}_2 = c_2$
c_1	127	58	92	14
c_2	35	62	131	194

Tabella 2.5: *esempio di attribuzione della classe ad un utente applicando il metodo BKS.*

Nell'esempio presentato in precedenza, se per un nuovo item si ricavasse la combinazione $\hat{S}_1 = c_1$ e $\hat{S}_2 = c_1$, esso sarebbe assegnato alla classe c_1, la quale risulta associata alla riga con il valore più elevato per la colonna corrispondente alla stima fornita dai classificatori di base (vedi Tabella 2.5).

É anche possibile stabilire una soglia oltre la quale ritenere affidabili i risultati: ad esempio, nel caso della configurazione delle stime $\hat{S}_1 = c_1$ e

$\hat{S}_2 = c_2$ la differenza tra i valori associati alle due classi è poco significativa, e potrebbe dare origini ad errori nella classificazione.

2.4.3 Algoritmi basati su meta-classificatori

Gli algoritmi di questo tipo utilizzano le stime fornite dai classificatori di base come meta-dati per un ulteriore livello di classificazione, in cui viene ricavato un modello per la caratterizzazione degli item in esame. Nel seguito sono descritti i principali algoritmi basati su questa metodologia operativa.

Stacked Generalization

L'algoritmo *Stacked-Generalization*, presentato da Wolpert in [62] è stato uno dei primi ad utilizzare il concetto di meta-dati per la classificazione di item. Ad un sottoinsieme degli item, di cui sia nota la classe di appartenenza, si applicano separatamente tutti i classificatori di base, con l'obiettivo di costruire un insieme di meta-apprendimento formato dalla reale classe di appartenenza dell'item a da tutte le stime fornite dai classificatori di base. A partire da tale insieme, un classificatore finale, che può essere uno dei classificatori di base già utilizzato in precedenza o un ulteriore algoritmo, elabora un modello da utilizzare per la classificazione degli item di cui non sia nota l'effettiva classe di appartenenza.

Combiner

Questa metodologia nasce come evoluzione della Stacked Generalization appena descritta. Come descritto da Chan e Stolfo in [10], in questo caso l'insieme di meta-apprendimento per generare il modello attraverso cui efettuare la classificazione finale degli utenti è costruito applicando una speciale regola detta *composition rule*, che consente di combinare tra di loro le stime fornite dai vari classificatori di base.
Esempi di composition rule sono i seguenti:

1. l'insieme di meta-apprendimento è costituito dall'etichetta rappresentante l'effettiva classe di appartenenza dell'item e dalle stime fornite dai classificatori di livello precedente. É possibile notare che, nel caso si utilizzi tale regola, questa metodologia è del tutto equivalente alla Stacked-Generalization descritta in precedenza;

2. l'insieme di meta-apprendimento è costituito dagli attributi associati all'item, dall'etichetta rappresentante l'effettiva classe di appartenenza dell'item e dalle stime fornite dai classificatori di livello precedente;

3. l'insieme di meta-apprendimento è costituito dagli attributi associati all'item, dall'etichetta rappresentante l'effettiva classe di appartenenza dell'item, dalle stime fornite dai classificatori di livello precedente e dall'accuratezza associata ad ogni stima;

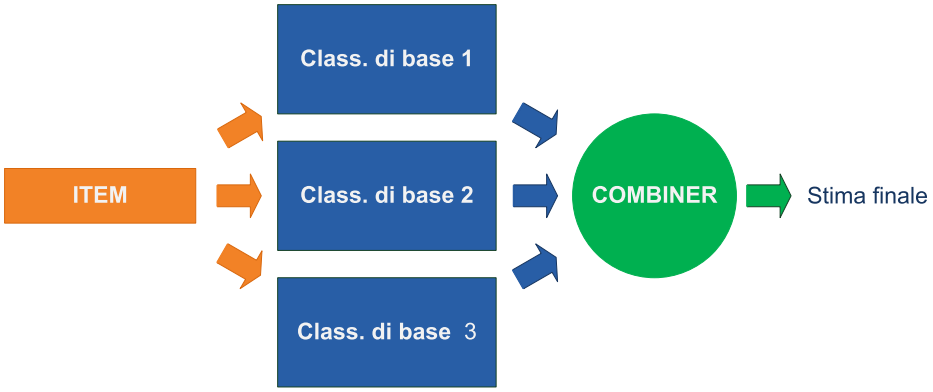


Figura 2.2: *uso della metodologia del Combiner per la stima della classe di un item. Le stime fornite dai tre classificatori di base sono utilizzate per costruire il meta-modello da cui il Combiner ricava la stima per la classe di appartenenza dell'item.*

L'insieme di meta-apprendimento viene in seguito utilizzato da un classificatore, detto *combiner*, che ricava da esso il modello per la classificazione degli item. Il ruolo di combiner può essere svolto da uno dei classificatori di base o da un ulteriore algoritmo.

Chan e Stolfo in [11] hanno sviluppato una soluzione basata sull'uso di più livelli di combiner, detta *Combiner Tree*, in cui si costruisce un albero le cui foglie sono rappresentate dai classificatori di base, mentre i livelli intermedi sono rappresentati da dei combiner. L'albero viene costruito secondo una metodologia bottom-up: le stime fornite da due o più classificatori di base sono utilizzate come meta-dati da un primo livello di combiner. Le stime ricavate da questi ultimi sono poi comunicate ai combiner di secondo livello, che le utilizzeranno a loro volta come meta-dati. Il procedimento viene iterato per un numero prestabilito di livelli, fino ad arrivare alla creazione della radice

dell'albero, responsabile della costruzione del modello per la classificazione finale degli item. L'uso di più livelli di combiner incrementa l'accuratezza delle stime finali, ma impone un incremento dei tempi di esecuzione, derivato dalla necessità di eseguire più volte il combiner e di elaborare i meta-dati per ogni livello dell'albero.

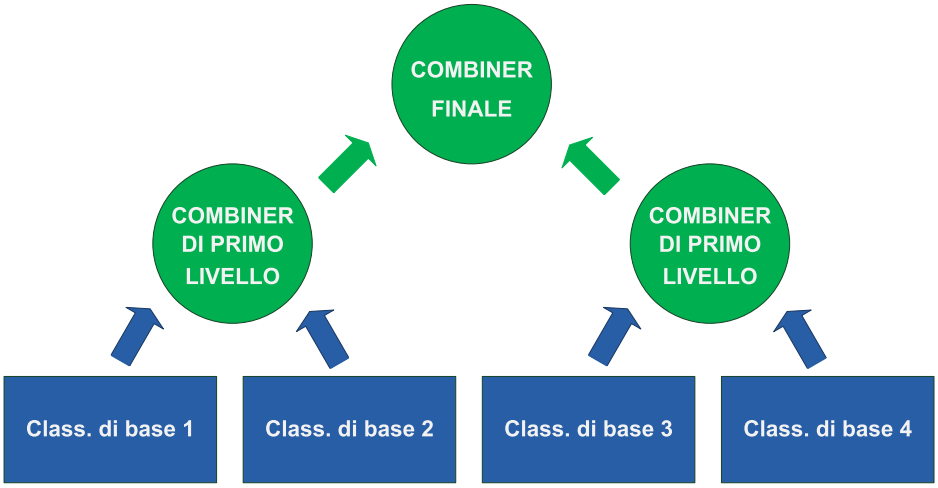


Figura 2.3: *la metodologia del Combiner Tree.*

Arbitro

Lo scopo di questa metodologia, presentata da Chan e Stolfo in [10], è costruire un classificatore in grado di predire la classe di appartenenza di item per cui vi siano contraddizioni nelle stime fornite dai classificatori di base.

Ciascun classificatore di base effettua separatamente la predizione della classe di appartenenza dell'item; se tutte le stime fornite sono concordi, l'item viene classificato come appartenente alla classe predetta. In caso contrario, invece, la stima dell'item sarà fornita da un ulteriore classificatore, detto *arbitro*, il quale opera su un insieme di meta-apprendimento costruito secondo una regola detta *Selection Rule*.
Esempi di Selection Rule sono i seguenti:

1. includere nell'insieme di meta-apprendimento gli attributi relativi agli item per cui non si verifica l'unanimità delle stime

2. includere nell'insieme di meta-apprendimento gli attributi relativi agli item per cui non si verifica l'unanimità delle stime e le classificazioni prodotte dai classificatori di base per tali item.

3. includere nell'insieme di meta-apprendimento tutti gli attributi relativi agli item per cui non si verifica l'unanimità delle stime, oltre agli attributi relativi agli item per cui i classificatori di base sono concordi nella predizione, che però risulta errata.

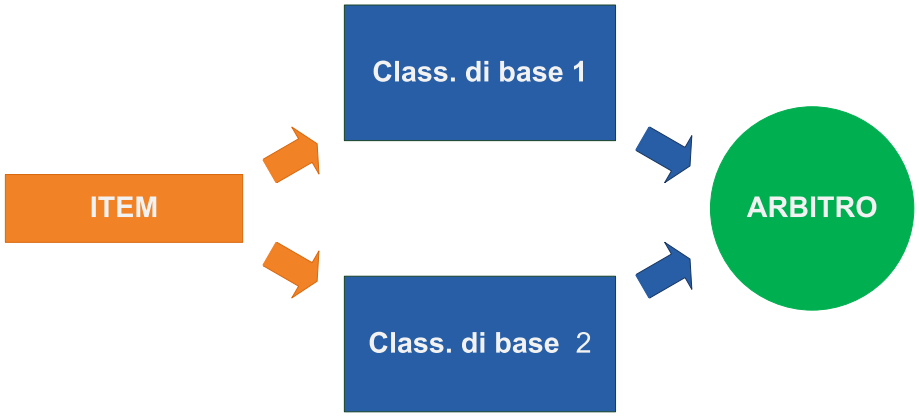


Figura 2.4: *costruzione dell'insieme di meta-apprendimento per l'arbitro a partire dalle stime fornite dai classificatori di base.*

L'arbitro ricava il modello per la predizione a partire dall'insieme di meta-apprendimento e assegna ad una classe di appartenenza a tutti quegli item per cui siano stati individuati conflitti nelle stime fornite in precedenza dai classificatori di base. L'item può quindi essere assegnato direttamente alla classe predetta dall'arbitro, oppure si può ricorrere alla metodologia del Voting, assegnando eventualmente un peso maggiore al voto fornito dall'arbitro rispetto a quelli assegnati ai voti dei classificatori di base.

Anche per questa metodologia è stata proposta una soluzione a più livelli [11], basata sulla creazione di un albero. I classificatori di base costituiscono le foglie dell'albero, esattamente come accade nella metodologia Combiner Tree, mentre i livelli intermedi sono in questo caso rappresentati da degli arbitri, i quali operano su insiemi di meta-apprendimento ricavati dalle stime fornite dal livello inferiore.

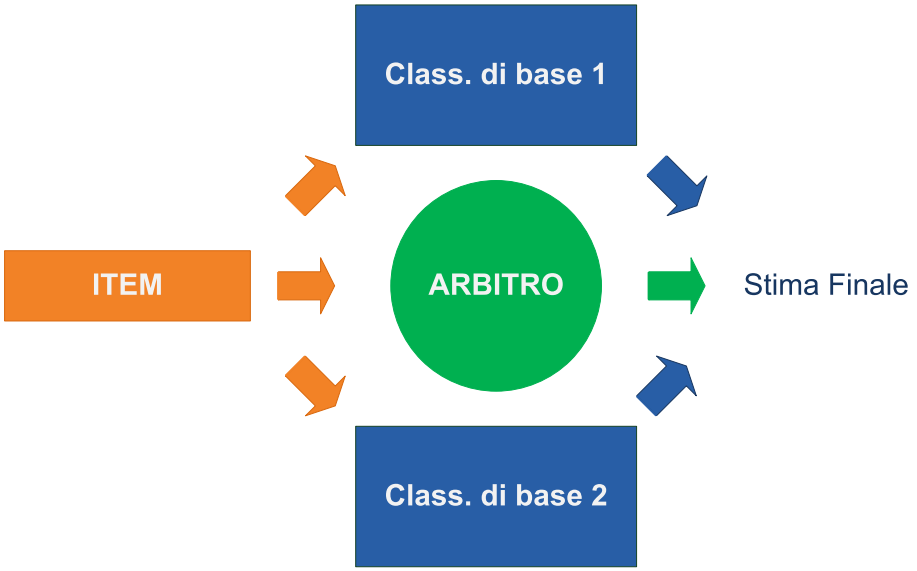


Figura 2.5: *assegnazione degli utenti alle classi; in questa fase si fa riferimento sia alle stime fornite dai classificatori di base che alla stima dell'arbitro.*

Grading

In questa metodologia, descritta da Seewald e Fürnkranz in [39], per ciascun classificatore di base utilizzato viene costruito un ulteriore classificatore, detto *grader*, il quale ha il compito di valutare l'attendibilità delle stime predette dal classificatore di base. Il grader può essere ricavato utilizzando il medesimo algoritmo del classificatore di base oppure, in alternativa, utilizzando un algoritmo differente.

In una prima fase si utilizza un insieme di apprendimento, formato da item di cui sia nota la classe di appartenenza, e si effettua un confronto tra la stima fornita dal classificatore di base e la classe a cui appartiene effettivamente l'item. Il risultato del confronto viene inserito in una etichetta, generalmente contenente il valore 1 se il confronto è positivo e -1 in caso di errore da parte del classificatore, la quale viene a sua volta concatenata agli attributi dell'item per formare l'insieme di meta-apprendimento. Il grader, a questo punto, ricava da tale insieme un modello per prevedere gli errori del classificatore di base, utilizzando l'etichetta contenente la valutazione della stima del classificatore di base.

Per gli item di cui non è nota la classe di appartenenza vengono quindi prodotte due stime: una da parte del classificatore di base, il cui risultato è la classe a cui assegnare l'item, l'altra, fornita dal grader a partire dal modello ricavato dall'insieme di meta-apprendimento, contenente la valutazione di tale stima. L'item viene effettivamente assegnato alla classe predetta dal classificatore di base solo in caso di valutazione positiva da parte del grader; in caso contrario la predizione non è considerata affidabile.

Capitolo 3

Dataset

Indice

In questo capitolo sono presentati i dataset MovieLens (Paragrafo 3.1) e Yahoo! (Paragrafo 3.2) utilizzati nell'ambito di questo lavoro di tesi. In particolare viene descritta la struttura delle singole basi di dati, soffermandosi sulle distribuzioni degli utenti tra le classi corrispondenti ai Lifestyle utilizzati, andando ad evidenziarne le peculiarità e le caratteristiche comuni.

3.1 Il dataset MovieLens

Il dataset MovieLens è stato sviluppato dal GroupLens Research Project presso il Dipartimento di Computer Science and Engineering dell'Università

del Minnesota [54]: trattandosi di un dataset liberamente accessibile è stato ampiamente utilizzato per progetti di ricerca nell'ambito dei sistemi di raccomandazione e dell'inferenza del Lifestyle degli utenti.

All'interno della base di dati sono contenute informazioni relative agli utenti, ai film da essi visionati e alle valutazioni espresse; attualmente esistono tre versioni del dataset, per ciascuna delle quali è garantito un numero minimo di valutazioni per ciascun utente pari a 20 film:

- la versione base contiene 943 utenti e 1 682 film, per un totale di circa centomila rating;
- una versione più ampia è formata da 6 040 utenti e 3 883 film, per un totale di circa un milione di valutazioni;
- la terza versione contiene 71 567 utenti e 10 681 film, per un totale di circa dieci milioni di rating.

Nel nostro lavoro abbiamo utilizzato la seconda versione del dataset, nella quale sono inserite esattamente 1 000 209 valutazioni; il numero medio di film valutati da ciascun utente è quindi pari a 165,6.
Lo schema ER per il dataset MovieLens è riportato nella Figura 3.1:

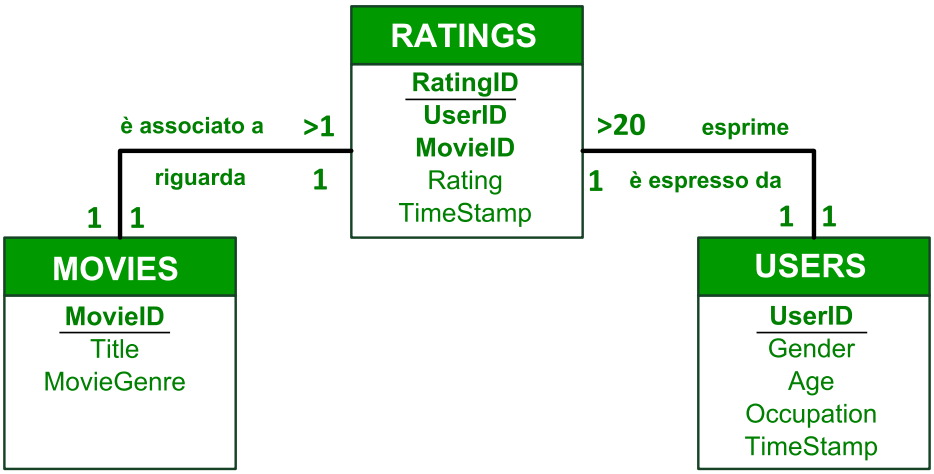


Figura 3.1: *lo schema ER per il dataset MovieLens.*

Utenti, film e valutazioni sono descritti da tre tabelle distinte. Come accennato in precedenza, un utente deve esprimere un numero di valutazioni maggiore

di 20, mentre per ognuno dei film presenti nel dataset deve essere presente almeno una valutazione da parte degli utenti. Un rating è associato solamente all'utente che ha effettuato la valutazione e fa riferimento ad un unico film.

Tabella USERS

Questa tabella contiene tutte le informazioni relative agli utenti; la struttura dei record è costituita dai seguenti campi:

- *UserID*: rappresenta la chiave primaria della tabella, utilizzata per identificare gli utenti univocamente. La chiave è totalmente anonima per evitare possibili violazioni alla privacy degli utenti.

- *Gender*: esprime il sesso dell'utente, indicato con le lettere 'M' per gli utenti di sesso maschile, ed 'F' per gli utenti di sesso femminile. La Tabella 3.1 mostra la ripartizione tra gli utenti maschi e femmine contenuti nel dataset.

CLASSE	Numero utenti	Percentuale utenti
Maschi	4331	71,71%
Femmine	1709	28,29%

Tabella 3.1: *ripartizione degli utenti del dataset MovieLens per SESSO.*

- *Age*: identifica la classe di età a cui appartiene l'utente; sono previste sette classi di età, la cui composizione è riportata nella Tabella 3.2.

- *Occupation*: indica la professione dell'utente; sono proposte 21 categorie di professioni. Tale informazione non rientra nel nostro ambito di applicazione e, pertanto, non viene considerata.

- *ZipCode*: rappresenta il codice postale dell'indirizzo dell'utente. Anche questa informazione non è stata considerata in questo lavoro.

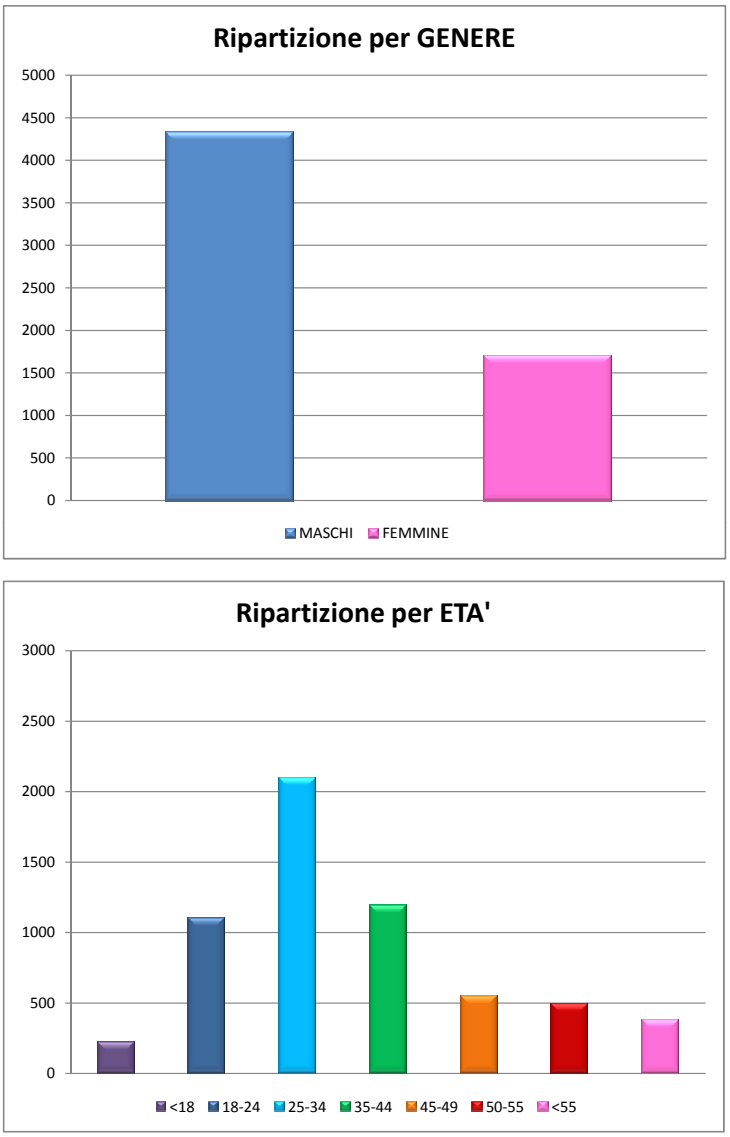


Figura 3.2: *ripartizione degli utenti del dataset MovieLens per SESSO ed ETA'.*

CLASSE	Numero utenti	Percentuale utenti
[<18]	222	3,68%
[18-24]	1103	18,26%
[25-34]	2096	34,70%
[35-44]	1193	19,75%
[45-49]	550	9,11%
[50-55]	496	8,21%
[>55]	380	6,29%

Tabella 3.2: *ripartizione degli utenti del dataset MovieLens per classi di ETA'.*

Tabella MOVIES

Per quanto riguarda i film contenuti nel dataset vengono fornite le seguenti informazioni:

- *MovieID*: è la chiave primaria necessaria all'identificazione univoca di ciascun film.
- *Title*: campo di testo che contiene il titolo del film.
- *MovieGenre*: indica il genere del film; nel dataset sono contenuti film appartenenti a 18 generi differenti, ad ognuno dei quali corrisponde un identificativo numerico. Ogni film può appartenere ad un massimo di sei generi. La Tabella 3.3 riporta l'elenco dei generi ed il numero di film appartenenti a ciascuno di essi.

GENERE	Numero di film	Percentuale sul totale
Animazione	105	1,64%
Film per bambini	251	3,92%
Commedia	1200	18,73%
Avventura	283	4,42%
Fantasy	68	1,06%
Romantico	471	7,35%
Drammatico	1603	25,02%
Azione	503	7,85%
Crime	211	3,29%
Thriller	492	7,68%
Horror	343	5,35%
Sci-fi	276	4,31%
Documentario	127	1,98%
Guerra	143	2,23%
Musical	114	1,78%
Mistery	106	1,65%
Noir	44	0,69%
Western	68	1,06%

Tabella 3.3: *ripartizione dei utenti del dataset MovieLens per GENERE.*

Tabella RATINGS

Contiene tutte le valutazioni espresse dagli utente, relativamente ai film disponibili nel dataset. I record appartenenti a questa tabella contengono i seguenti campi:

- *RatingID*: chiave primaria della tabella.

- *UserID*: identifica l'utente che ha espresso la valutazione. Rappresenta una chiave esterna, utilizzata per individuare l'utente associato al rating.

- *MovieID*: identifica il film valutato. Anche in questo caso si tratta di una chiave esterna.

- *Rating*: indica il voto assegnato dall'utente al film. Le valutazioni sono comprese tra un valore pari a a 1, che indica il mancato gradimento dell'utente, ed un valore pari a 5, rappresentativo del massimo livello di apprezzamento.

- *TimeStamp*: fornisce un riferimento temporale per la valutazione. Questo tipo di informazione non rientra nel nostro ambito applicativo e non viene di conseguenza considerata.

3.2 Il dataset Yahoo!

Questo dataset è stato realizzato nell'ambito del programma Yahoo! Research Alliance [58] con l'intento di sviluppare una base di dati utilizzabile all'interno di progetti di ricerca e sviluppo di soluzioni basate su algoritmi di raccomandazione.

All'interno del dataset sono contenuti dati relativi a 11 915 film unitamente ad informazioni demografiche riguardanti 7 642 utenti, i quali hanno espresso un totale di 211 231 valutazioni. Tutti gli utenti inseriti nel dataset hanno generato almeno 10 valutazioni ciascuno e per ognuno dei film è presente almeno un rating. Il numero medio di valutazioni espresse è pari a 27,64 per gli utenti e a 17,73 per i film.

La struttura della base di dati, mostrata nella Figura 3.3, è molto simile a quella del dataset MovieLens descritta in precedenza. Le tre tabelle principali contengono rispettivamente le informazioni relative agli utenti, ai film visualizzati e alle valutazioni fornite.

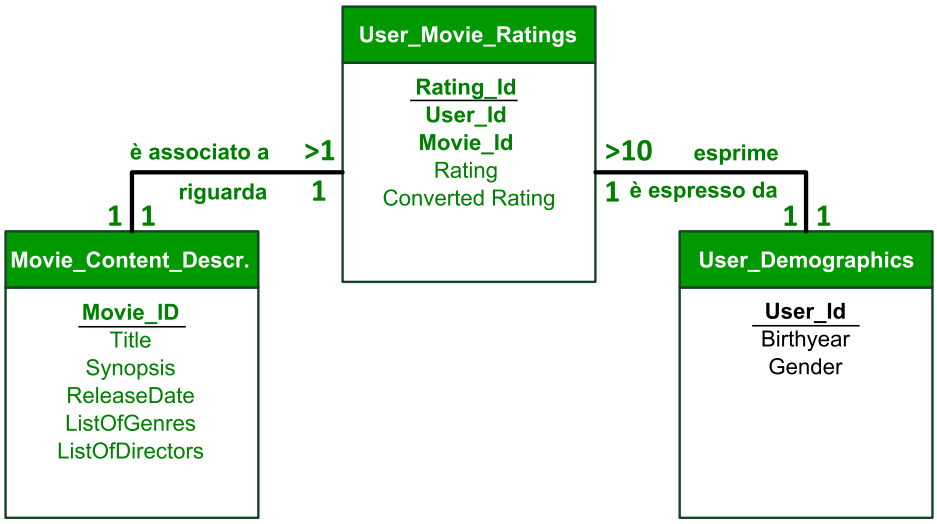


Figura 3.3: *lo schema ER per il dataset Yahoo!.*

Tabella User_Demographics

I record per la tabella relativa agli utenti contengono tre campi: *User_ID*, *Gender* e *Birthyear*.

- *User_ID*: chiave primaria necessaria all'identificazione univoca degli utenti. Anche in questo caso per garantire l'anonimato la chiave è rappresentata da un identificativo numerico.

- *Gender*: rappresenta il sesso dell'utente, indicato dai caratteri 'm' ed 'f'. Nel dataset sono presenti 23 utenti per cui questo valore non è indicato, che non sono stati quindi presi in considerazione nel nostro studio. La ripartizione degli utenti tra maschi e femmine è mostrata nella Tabella 3.4.

CLASSE	Numero utenti	Percentuale utenti
Maschi	5436	71,13%
Femmine	2183	28, 57%
Non Assegnati	23	0, 3%

Tabella 3.4: *ripartizione degli utenti del dataset Yahoo! per SESSO.*

- *Birthyear*: indica l'anno di nascita dell'utente; la Tabella 3.5 mostra la ripartizione degli utenti in classi a seconda di tale informazione demografica.

CLASSE	Numero utenti	Percentuale utenti
[1986-2001]	1776	23,24%
[1980-1985]	2577	33,72%
[1975-1979]	1096	14,34%
[1970-1974]	737	9,64%
[1960-1969]	853	11,16%
[1950-1959]	371	4,85%
[fino al 1949]	232	3,04%

Tabella 3.5: *ripartizione degli utenti del dataset Yahoo! per ANNO DI NASCITA.*

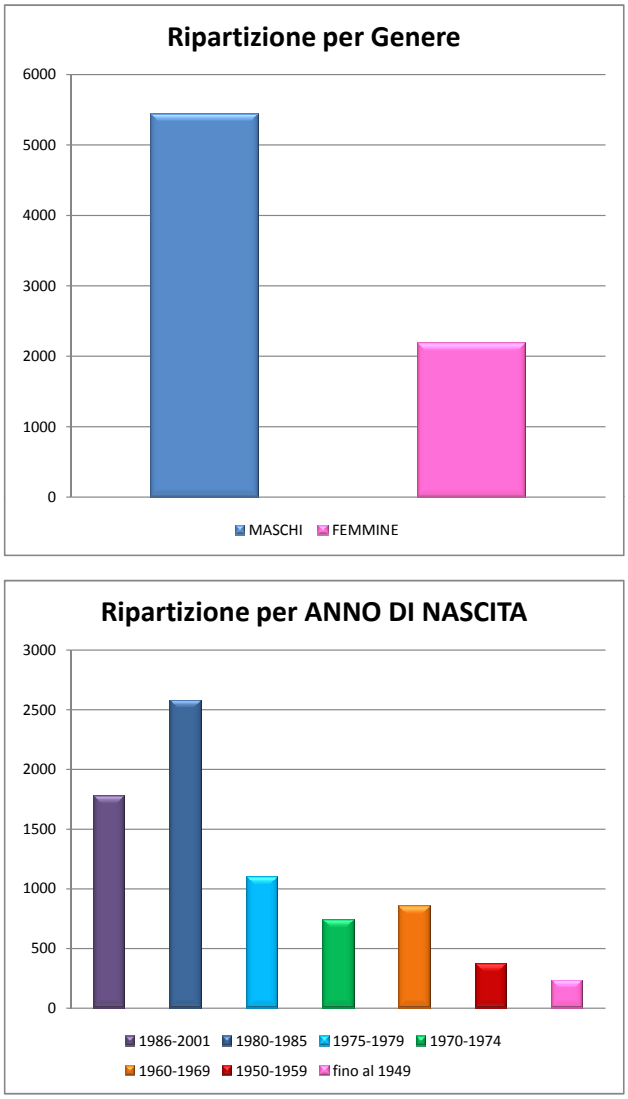


Figura 3.4: *ripartizione degli utenti del dataset Yahoo! per SESSO ed ANNO DI NASCITA.*

Tabella Movie_Content_Description

Per quanto riguarda i film contenuti nel dataset viene fornita una lista molto dettagliata di informazioni, tra le quali citiamo le più significative:

- *Movie_ID*: che rappresenta la chiave primaria della tabella;
- titolo del film;
- breve descrizione della trama;
- lista dei generi a cui appartiene il film (vedi Tabella 3.6 per l'elenco dei 18 possibili generi a cui può essere associato un film);
- nome ed identificativo del regista;
- nome ed identificativo degli attori principali;
- data di uscita del film;
- principali premi vinti.

Tabella User_Movie_Ratings

Relativamente alle valutazioni espresse sono invece inserite nel dataset le seguenti informazioni:

- *Rating_Id*: rappresenta la chiave per identificare le valutazioni.
- *User_Id*: chiave esterna che identifica l'utente che ha espresso la valutazione.
- *Movie_Id*: chiave esterna per associare la valutazione al film corrispondente.
- *Rating*: valutazione espressa dall'utente; viene proposta una scala di voti in tredici livelli, in cui il livello 1 indica il livello di apprezzamento più basso da parte dell'utente.
- *Converted Rating*: valutazione convertita secondo una scala a cinque livelli, analoga a quella utilizzata nel dataset MovieLens.

GENERE	Numero di film	Percentuale sul totale
Special Interest	457	0,13%
Western	1808	0,53%
Azione/Avventura	65108	18,95%
Suspence/Horror	13974	4,07%
Film per bambini	17471	5,09%
Commedia	69028	20,11%
Documentario	975	0,28%
Thriller	30884	9,00%
Drammatico	51523	15,01%
Arte	5563	1,62%
Crime/Gangster	24636	7,18%
Romantico	18542	5,40%
Science Fiction	30654	8,93%
Musical	5443	1,59%
Animazione	7061	2,06%
Vari	11	0,01%
Film per adulti	13	0,01%
Reality	49	0,01%

Tabella 3.6: *ripartizione dei utenti del dataset MovieLens per GENERE.*

Capitolo 4

Framework

Indice

Questo capitolo è dedicato alla definizione del framework sviluppato; in particolare nel Paragrafo 4.1 sono proposte alcune metriche di valutazione necessarie per la misurazione dei risultati ottenuti nei singoli test eseguiti, mentre il Paragrafo 4.2 contiene la descrizione dello schema implementativo comune utilizzato per effettuare i test relativi alla classificazione degli utenti applicando gli algoritmi di base presentati nel Capitolo 2.

4.1 Metodologia di valutazione

Per la valutazione della qualità dei vari metodi presi in considerazione sono state individuate delle metriche comuni da utilizzare sia nella fase di ricerca della configurazione di parametri che assicuri i risultati migliori per ciascun metodo, sia nella fase di valutazione globale e confronto tra i risultati ottenuti nelle differenti tipologie di test.

4.1.1 Metriche di valutazioni tradizionali

Come accennato in precedenza esiste un forte legame tra il Targeted Advertising ed i sistemi di raccomandazione; alcune delle metriche originariamente

definite per questo ambito applicativo possono quindi essere opportunamente adattate, in modo da consentire di effettuare la valutazione di soluzioni finalizzate alla profilazione degli utenti.

I sistemi di raccomandazione rappresentano un settore in fase di forte crescita, grazie alla loro versatilità che consente di applicarli in un gran numero di contesti (libri, film, musica, ecc.); l'utilizzo di sistemi di raccomandazione, inoltre, introduce notevoli benefici sia per i clienti, sia per i fornitori di servizi, i quali sono in grado di proporre i prodotti che soddisfano maggiormente i gusti degli utenti, e che quindi hanno maggiori probabilità di essere acquistati. In particolare, un sistema di questo tipo si occupa di proporre all'utente una lista di item che dovrebbero soddisfare i propri gusti e preferenze basandosi, ad esempio, sull'analisi di feedback espressi o di comportamenti passati. La valutazione delle performance ottenute da un sistema di raccomandazione avviene individuando le seguenti informazioni:

- True positive (TP): numero di item di interesse raccomandati all'utente;
- True negative (TN): numero di elementi non di interesse che non vengono proposti all'utente;
- False negative (FN): numero di item di interesse per l'utente che non vengono raccomandati;
- False positive (FP): numero di elementi non di interesse per l'utente che vengono comunque suggeriti;.

A partire da queste informazioni è quindi possibile effettuare il calcolo di alcune metriche di valutazione, tra le quali le più significative sono la *Precision*, l'*Accuracy* e la *Recall* [3].

Precision e Accuracy

Nell'ambito dei sistemi di raccomandazione la *Precision* è definita come il numero di item rilevanti raccomandati, rispetto al totale delle raccomandazioni effettuate; la formula per il calcolo di questa metrica è quindi la seguente:

$$\text{Precision} = \frac{TP}{TP + FP}$$

L'*Accuracy*, invece, è calcolata tenendo conto anche dei cosiddetti True Negative e False Negative, che rappresentano rispettivamente il numero di item che vengono correttamente riconosciuti come di non interesse per l'utente e quindi non rientrano nella lista delle raccomandazioni proposte, e il numero di item di interesse che non vengono suggeriti all'utente.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FN + FP}$$

La differenza fondamentale tra le due metriche consiste nel fatto che, mentre la Precision si limita a considerare gli item suggeriti all'utente, valutando la percentuale di raccomandazioni corrette sul totale effettuato, il calcolo dell'Accuracy comprende anche la valutazione delle performance relative agli item che non fanno parte della lista di raccomandazioni proposta all'utente.

La Precision può essere adattata all'utilizzo nell'ambito del Targeted Advertising per la valutazione delle prestazioni di problemi binari nei quali soltanto una delle due classi sia considerata rilevante e, di conseguenza, le uniche performance da valutare siano quelle relative a tale classe. In particolare l'uso di questa metrica consente di calcolare la percentuale di utenti assegnati correttamente rispetto al totale delle assegnazioni effettuate relativamente alla classe rilevante.

Nel nostro lavoro ci troviamo invece di fronte a problemi binari in cui entrambe le classi sono considerate rilevanti, come ad esempio quelli relativi all'inferenza del sesso degli utenti, oppure a problemi che prevedono un numero di classi superiore a due, come nel caso della classificazione degli utenti in base all'età. Si è scelto di conseguenza di non utilizzare la Precision, ma di ricorrere all'uso dell'Accuracy, opportunamente adattata per le esigenze del nostro ambito applicativo.

Relativamente all'ambito del Targeted Advertising, infatti, l'Accuracy globale può essere definita come il rapporto tra il numero di utenti che sono stati assegnati correttamente alla propria classe di appartenenza, ed il totale degli elementi per i quali è stata effettuata l'assegnazione ad una classe.

$$\text{Accuracy}_{globale} = \frac{\text{\# utenti correttamente classificati}}{\text{\# utenti assegnati ad una classe}}$$

Un valore di Accuracy pari ad 1 indica che tutte le previsioni fornite dall'algoritmo relativamente alla classe di appartenenza degli utenti presenti nell'insieme di test sono esatte; il raggiungimento di un livello inferiore di

Accuracy, invece, indica la presenza di errori nella fase di classificazione. Oltre che a livello globale, è possibile calcolare l'Accuracy anche per le singole classi, limitandosi a considerare il numero di utenti correttamente assegnati ed il totale degli utenti assegnati relativamente alla classe in esame.

Recall

Per quanto riguarda la *Recall*, invece, la formula utilizzata nell'ambito dei sistemi di raccomandazione è la seguente:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Anche questa metrica può essere adattata al nostro scenario applicativo, definendola come la percentuale degli utenti per i quali è stata fornita una classificazione corretta, rispetto al totale degli elementi presenti nell'insieme di test. Analizzando tale indicatore, quindi, è possibile stabilire se il metodo sia in grado o meno di individuare un numero soddisfacente di utenti.
La formula per il calcolo della Recall globale relativamente agli algoritmi da noi proposti è la seguente:

$$\text{Recall}_{globale} = \frac{\text{\# utenti correttamente classificati}}{\text{totale utenti}}$$

Un valore di Recall pari ad 1 indica che l'algoritmo è stato in grado di individuare tutti gli utenti presenti nell'insieme di test, mentre valori più bassi indicano il fatto che per alcuni degli utenti non è stata effettuata una previsione sulla classe di appartenenza o tale previsione è risultata errata.

Alcuni algoritmi impongono che tutti gli utenti analizzati siano assegnati ad una classe, senza lasciare la possibilità che vi siano elementi per i quali non venga effettuato l'assegnamento: in questo caso i valori di Recall ed Accuracy globali coincidono, in quanto il numero di utenti per i quali viene effettuata la previsione è uguale al numero totale di elementi presenti nell'insieme analizzato.

Nella maggioranza delle situazioni, ad un aumento di una delle due metriche appena presentate corrisponde una diminuzione dell'altra. Per ottenere valori di Accuracy più elevati, infatti, è necessario utilizzare criteri maggiormente restrittivi nella fase di caratterizzazione degli utenti, ottenendo di conseguenza una diminuzione della Recall complessiva, dovuta alla mancata

assegnazione di alcuni utenti alle classi. Viceversa, il tentativo di incrementare il livello di Recall porta all'utilizzo di criteri meno selettivi, che rischiano di introdurre errori nella classificazione e una conseguente diminuzione dell'Accuracy complessiva.

Come per l'Accuracy, anche la Recall può essere calcolata per ciascuna delle singole classi previste, limitandosi a considerare il totale degli utenti che appartengono alla classe in questione ed il numero di successi per essa ottenuti.

4.1.2 Lift

In molti settori la selezione delle informazioni pubblicitarie da proporre agli utenti si basa sull'analisi di rilevazioni statistiche effettuate su campioni di utenti che fruiscono del servizio; uno degli esempi più significativi è rappresentato dal settore televisivo, in cui gli operatori dispongono delle rilevazioni elaborate da Auditel [50], che consentono di conoscere per ciascun programma trasmesso la composizione del pubblico, suddividendo gli utenti in un insieme di categorie target prestabilite.

Le usuali strategie di marketing applicate prevedono di individuare il target più rilevante all'interno del pubblico complessivo del programma, proponendo di conseguenza le informazioni pubblicitarie di maggiore interesse per tale categoria. Lo svantaggio di questo tipo di strategia consiste nel fatto che i medesimi messaggi pubblicitari sono proposti a tutti gli utenti, anche a coloro i quali possono avere interessi e preferenze significativamente differenti rispetto a quelli espressi dal target individuato. Ciò si traduce in un minore coinvolgimento di questi utenti rispetto alle campagne pubblicitarie proposte ed in una diminuzione della redditività da esse prodotta. Nonostante la presenza di tali svantaggi, il ricorso a questa strategia di marketing è giustificato dall'impossibilità di fornire informazioni pubblicitarie personalizzate agli utenti attraverso il canale televisivo tradizionale.

Per settori nei quali è invece possibile una elevata personalizzazione delle informazioni pubblicitarie, come ad esempio l'IPTV, gli svantaggi che si verificano ricorrendo alle strategie tradizionali possono essere eliminati ricorrendo a tecniche di Targeted Advertising, nelle quali le informazioni pubblicitarie differiscono a seconda del profilo individuato per ciascun utente.

Tuttavia, perché questo tipo di soluzioni sia effettivamente più efficace rispetto alle strategie tradizionali è necessario che la percentuale degli utenti

per i quali viene effettuata una classificazione corretta sia superiore rispetto alla percentuale di utenti appartenenti alla classe maggiormente popolata tra quelle previste. In caso contrario, infatti, il numero di utenti per i quali si andrebbero a proporre informazioni pubblicitarie effettivamente rilevanti applicando il Targeted Advertising sarebbe inferiore rispetto a quello che si otterrebbe proponendo lo stesso messaggio a tutti gli utenti, come avviene nelle campagne pubblicitarie non personalizzate.

Utilizzando le metriche di valutazioni tradizionali non è possibile effettuare un'analisi di questo tipo: sorge quindi la necessità di definire una nuova metrica che sia in grado di valutare le performance della fase di classificazione degli algoritmi utilizzati nelle soluzioni Targeted Advertising, confrontandole con quelle che si otterrebbero assegnando tutti gli utenti alla classe maggiormente popolata tra quelle a disposizione.
In [39] viene utilizzata una metrica chiamata *Lift*, calcolata come:

$$\text{Lift}_{cl} = \frac{\text{Accuracy}_c}{\text{P}_{cl}}$$

in cui il valore di P_{cl} indica la probabilità che un utente u scelto a caso appartenga alla classe cl, calcolata come rapporto tra il numero di utenti appartenenti alla classe ed il totale degli utenti dell'insieme preso in considerazione.

Per giustificare il ricorso alla metodologia oggetto della valutazione è necessario l'ottenimento di un valore di Lift superiore ad uno, il quale indica che, per la classe in esame, il numero di utenti raggiunto da un messaggio pubblicitario pertinente è superiore applicando la metodologia in esame piuttosto che ricorrendo alle tecniche tradizionali.

L'utilizzo di tale metrica, tuttavia, consente di ottenere informazioni relativamente alle performance ottenute rispetto alle singole classi, ma non di effettuare valutazioni complessive sul metodo, estese a tutte le classi in esame; per superare questa limitazione viene nel seguito proposta una nuova metrica di valutazione, in grado di confrontare efficacemente le prestazioni complessive di una metodologia basata sul Targeted Advertising con quelle ottenibili senza applicare tale tecnica.

$$\text{Lift} = \frac{\text{performance complessive ottenute utilizzando TA}}{\text{perfomance complessive ottenute con soluzioni tradizionali}}$$

Le performance ottenute applicando la metodologia di Targeted Advertising oggetto della valutazione sono espresse moltiplicando il valore di Accuracy relativo a ciascuna classe per il numero di utenti che sono stati effettivamente assegnati dal metodo a tale classe. Sommando i risultati ottenuti per ognuna delle classi si ottiene il numero complessivo di utenti per i quali il metodo in esame ha fornito una stima corretta.

Moltiplicando il numero totale degli utenti da classificare per la percentuale di utenti appartenenti alla classe più popolata tra quelle previste si ricava, invece, il numero di utenti appartenenti a tale classe, che rappresenta un indicatore per le performance ottenibili senza ricorrere al Targeted Advertising.

La formula per il calcolo della metrica proposta è quindi la seguente:

$$\text{Lift} = \frac{\sum_{c=1}^{\#\text{ classi}} (\text{Accuracy}_{cl} * \text{utenti assegnati a } c)}{\text{totale utenti} * max(\text{P}_c)} \tag{4.1}$$

Alcuni degli algoritmi di classificazione esistenti, tuttavia, prevedono la possibilità di non assegnare un utente ad alcuna delle classi previste, quando si verifichino situazioni di incertezza o di mancanza delle informazioni necessarie per effettuare una classificazione sufficientemente affidabile. Nelle applicazioni reali tali utenti vengono solitamente assegnati alla classe maggiormente popolata, in modo da minimizzare gli errori di classificazione.

Per tenere in considerazione anche questo aspetto è quindi necessario modificare la formula per il calcolo del Lift, introducendo un ulteriore termine a denominatore, che viene ricavato moltiplicando il numero di utenti non assegnati per la percentuale di utenti appartenenti alla classe più popolata.

La formula (4.1) viene di conseguenza modificata, ottenendo la seguente:

$$\text{Lift} = \frac{\text{tot. utenti non ass.} * max(\text{P}_c) + \sum_{c=1}^{\#\text{ classi}} (\text{Accuracy}_{cl} * \text{utenti assegnati a } c)}{\text{totale utenti} * max(\text{P}_c)}$$

L'ottenimento di un valore di Lift maggiore di uno indica che, applicando la metodologia di profilazione in esame, si ottiene una classificazione sufficientemente precisa ed accurata, la quale consente quindi di ricavare maggiori benefici rispetto alle tradizionali strategie di marketing, fornendo al contempo una solida giustificazione per il ricorso alle tecniche di Targeted Advertising.

L'utilizzo di questa nuova metrica di valutazione risulta quindi molto utile nella fase di pianificazione delle strategie di marketing da applicare, in quanto consente di effettuare una valutazione completa ed approfondita delle prestazioni ottenibili ricorrendo al Targeted Advertising, integrando le informazioni ricavate attraverso l'analisi delle metriche tradizionali presentate in precedenza.

4.2 Schema implementativo

In questa sezione è presentato lo schema implementativo generale utilizzato, a partire dalla costruzione delle matrici da analizzare fino alla realizzazione dei test per valutare l'efficacia di ciascun algoritmo. La metodologia operativa utilizzata per la categorizzazione degli utenti è la medesima per ciascuna delle soluzioni prese in considerazione, salvo alcune differenze imposte dal tipo di tecnica scelta e dagli strumenti necessari alla sua applicazione.

4.2.1 Costruzione delle matrici UGM

Come già chiarito in precedenza, lo scopo del nostro lavoro è individuare il Lifestyle di un utente, analizzando le valutazioni da esso espresse in merito ad un insieme di generi di film.

Le metodologie di analisi sono quindi state applicate a matrici User-Genre, costruite a partire dalle informazioni presenti nei dataset descritti nel Capitolo 3, in cui ogni riga corrisponde ad un utente, mentre ciascuna colonna è associata ad uno dei generi di film disponibili.

Più formalmente, definiti U il numero di record presenti nel dataset e G il numero di generi, UGM è una matrice UxG, in cui l'elemento $x_{(u,g)}$ indica la somma delle valutazioni espresse dall'utente u relativamente a film di genere g.

Per costruire le matrici UGM da utilizzare è quindi necessario individuare per ciascun utente tutti i rating espressi e, in seguito, sommare tra di loro tutte le valutazioni riguardanti film dello stesso genere; eseguendo queste operazioni si ricava un vettore riga rappresentante le preferenze espresse dall'utente, che sarà formato da tante colonne quanti sono i generi presenti nel dataset. Gli elementi di tali vettori assumono un valore pari a zero nel caso

l'utente non abbia espresso alcuna valutazione relativamente a film del genere associato all'elemento in questione.

Concatenando tra di loro i vettori riga associati a ciascun utente si ricavano le matrici UGM, su cui applicare in seguito i metodi per effettuare l'inferenza dei Lifestyle degli utenti:

- la matrice UGM_{ML}, relativa al dataset MovieLens, formata da un numero di righe pari a 6040 e da un numero di colonne pari a 18;
- la matrice UGM_Y, che rappresenta il dataset Yahoo!, costituita da 7642 righe e 18 colonne.

Un esempio di vettore rappresentante le preferenze di un utente per il dataset MovieLens è il seguente, formato da 18 elementi, tanti quanti sono i generi di film contenuti nel dataset:

$$u_i = \begin{bmatrix} 0 & 0 & 12 & 4 & 47 & 9 & 0 & 13 & 54 & 10 & 0 & 6 & 31 & 7 & 2 & 6 & 19 & 0 \end{bmatrix}$$

I motivi per cui utilizziamo una matrice relativa agli utenti e ai generi sono legati ad una serie di benefici che si possono ottenere rispetto all'eventualità di considerare una struttura dati alternativa che prende il nome di *User Rating Matrix (URM)*, contenente la valutazioni espresse da ciascun utente relativamente ai film presenti nella base di dati, senza raggrupparli in base al genere. Una prima dimostrazione di questo fatto è contenuta all'interno del lavoro di Granara [19], dove gli stessi test eseguiti mostrano risultati migliori nel caso in cui si faccia riferimento ad una matrice di tipo UGM rispetto che ad una di tipo URM. Allo stesso modo, uno dei principali vantaggi dell'utilizzo di una matrice UGM è di tipo computazionale: la matrice UGM, nel nostro caso, è infatti composta solamente da 18 colonne, tante quanti sono i generi presenti in entrambi i dataset. La matrice URM, invece, è costituita da un numero di colonne molto più elevato, pari ad alcune migliaia, la cui gestione risulterebbe molto più gravosa. L'UGM si dimostra poi di notevole utilità nello sviluppo della tecnica del Porting descritta nel Capitolo 8, per cui si configura come l'unica struttura valida di riferimento, fornendo un'intelaiatura comune basata sui generi e consentendo un adattamento che, nel caso di una URM, sarebbe molto più complicato e comporterebbe la necessità di valutare dataset contenenti esattamente gli stessi film. A tutto questo si aggiunge infine il problema della corretta determinazione del Lifestyle degli utenti, che risulta

strettamente connesso ai generi dei film, come già provato da ricerche e studi precedenti [35, 36, 42]: utenti appartenenti a Lifestyle differenti hanno infatti preferenze non comuni relativamente ai generi di film preferiti.

4.2.2 Operazioni preliminari sui dataset

Dall'analisi delle matrici UGM ricavate dai dataset Yahoo! e MovieLens, è emerso come gli utenti abbiano espresso un maggior numero di valutazioni relativamente ai generi *"Azione/Avventura"*, *"Commedia"* e *"Drammatico"* che, di conseguenza, risultano essere i preferiti per un numero significativo di utenti. É emersa quindi la necessità di effettuare una *normalizzazione* delle matrici UGM, in modo da ottenere una distribuzione più omogenea delle valutazioni espresse dagli utenti per i vari generi presenti nei dataset.

La normalizzazione è stata effettuata dividendo ciascun valore $UGM_{(u,g)}$ contenuto nelle matrici per un fattore calcolato come:

$$normFactor_{(u,g)} = \sum(\text{film di genere } g \text{ valutati da } u)^{\omega}$$

Il valore assegnato al parametro ω può variare da zero ad uno. Nel primo caso l'effetto della normalizzazione risulterebbe nullo mentre, qualora ω assumesse il valore 1, la (4.2.2) restituirebbe il numero di rating espressi dall'utente u per film di genere g. Dividendo ciascun valore $UGM_{(u,g)}$, per il risultato così ottenuto, si ricaverebbe la valutazione media fornita da u relativamente al genere g.

Per scegliere il valore da assegnare a tale parametro sono stati quindi eseguiti dei test sull'inferenza del sesso degli utenti, applicando uno degli algoritmi esistenti, SVD (vedi Sezione Z) e valutando i risultati ottenuti in corrispondenza dei possibili valori assegnabili ad ω. Per effettuare tali sperimentazioni è stata utilizzata la metodologia della cross validation.

Dall'analisi dei risultati in termini di Recall, riportati in Tabella 4.1, è possibile notare come le migliori performance per il dataset Yahoo si ottengono con ω pari a $0{,}8$. Relativamente a Movielens, invece, il valore di tale parametro che garantisce i migliori risultati è pari a 0, seguito da $0{,}8$. Nei test analoghi che sono stati condotti usando le AR, la scelta del parametro si è mostrata non influenzare particolarmente la qualità del classificatore. A partire da tali considerazioni è stato quindi scelto di assegnare ad ω un valore pari a $0{,}8$, che è stato utilizzato in tutti i test eseguiti.

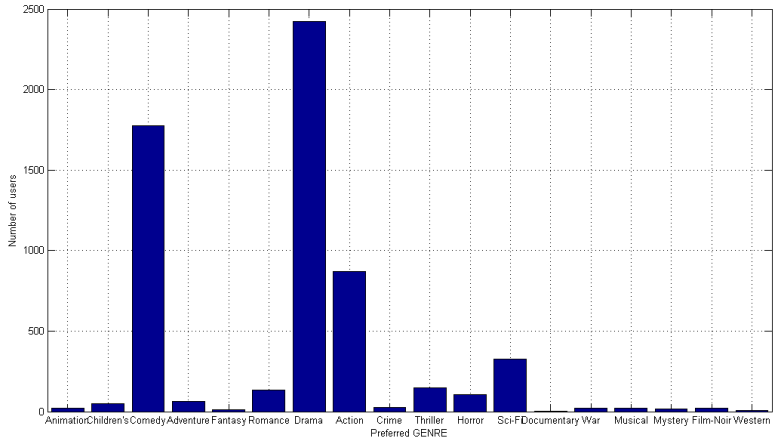
ω	Recall Yahoo!	Recall MovieLens
0,0	0,7139	0,7677
0,1	0,7161	0,7472
0,2	0,7183	0,7136
0,3	0,7205	0,7268
0,4	0,7301	0,7434
0,5	0,7244	0,7505
0,6	0,7155	0,7505
0,7	0,7196	0,7456
0,8	0,7309	0,7533
0,9	0,7235	0,7456
1,0	0,7287	0,7489

Tabella 4.1: *risultati dei test relativi all'inferenza sui sessi degli utenti utilizzando valori diversi del parametro ω. Le migliori performance in termini di Recall si ottengono per ω uguale a $0{,}8$. Tali risultati sono stati ricavati applicando SVD ad entrambi i dataset, ricorrendo alla tecnica della cross validation.*

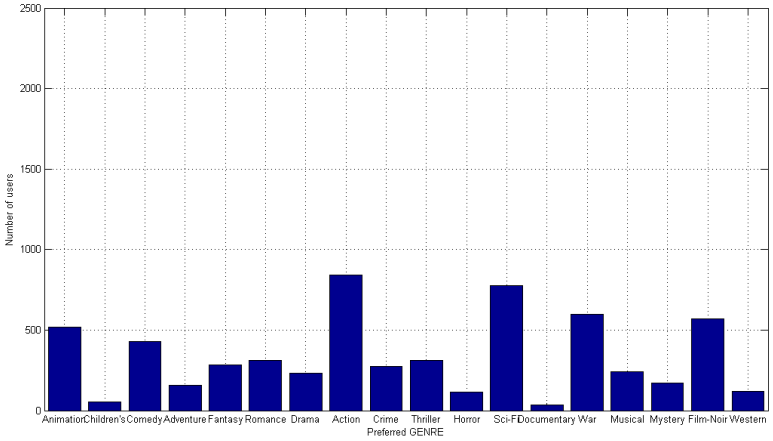
Le matrici ottenute applicando la normalizzazione sono state denominate UGMNorm; gli effetti di tale procedura sono mostrati nella Figura 4.1 in cui, facendo riferimento al dataset MovieLens, è riportata per ciascun genere la somma dei rating relativi a film che vi appartengono, prima e dopo la normalizzazione della matrice UGM.

Estrazione dei generi preferiti per ciascun utente

I lavori di Bianchi [7] e Granara [19] dimostrano come i migliori risultati relativamente all'applicazione di AR e SVD per la classificazione degli individui si ottengano considerano un numero limitato di generi di film per ciascun utente. A partire da questa considerazione, quindi, è stata inserita nello schema implementativo una fase di determinazione dei generi preferiti, in cui sono stati eliminati dalla matrice UGM tutti i valori associati a generi di film poco significativi per l'utente. Per determinare i generi di maggiore rilevanza per l'utente si è considerata la somma dei ratings espressi per ciascun genere, mantenendo soltanto i valori più elevati. Attraverso questa operazione si è scelto di concentrarsi soltanto su un numero limitato di generi, in modo da evitare che i risultati fossero influenzati dalla presenza di generi



(a) *Dataset MovieLens non normalizzato.*



(b) *Dataset MovieLens dopo la normalizzazione.*

Figura 4.1: *gli effetti della normalizzazione sul dataset MovieLens.*

poco significativi per l'utente.

4.2.3 Partizionamento del dataset

La prima operazione da effettuare è la suddivisione degli utenti a disposizione in due sottoinsiemi, rispettivamente detti di *apprendimento* e di *test* (vedi Figura 4.2). L'insieme di apprendimento è costituito da utenti di cui sono considerate note sia le preferenze espresse relativamente ai generi di film analizzati, sia le informazioni anagrafiche, in particolare l'età ed il sesso. Per quanto riguarda l'insieme di test, invece, si suppone di essere a conoscenza soltanto delle preferenze espresse dagli utenti, mentre le informazioni anagrafiche dovranno essere inferite tramite l'applicazione degli algoritmi selezionati.

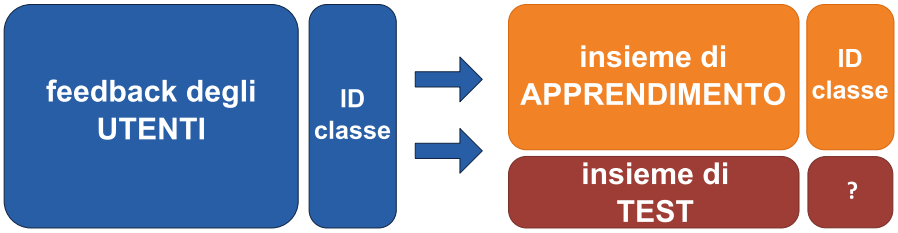


Figura 4.2: *la suddivisione dei record negli insieme di apprendimento e di test.*

4.2.4 Apprendimento ed applicazione del modello

Dopo aver partizionato l'insieme degli utenti, lo schema implementativo prevede la costruzione di un modello per la classificazione degli utenti dell'insieme di test, ricavato dall'applicazione degli algoritmi di data mining all'insieme di apprendimento. In seguito, per ciascuna delle metodologie da implementare, è necessario individuare quali parametri utilizzare per ottenere le migliori prestazioni nella fase di caratterizzazione degli utenti di test.

L'individuazione dei migliori parametri si realizza ricorrendo alla *n-fold Cross Validation*: questa tecnica prevede di effettuare un numero n di iterazioni, per ognuna delle quali si suddivide l'insieme di apprendimento iniziale in n partizioni, utilizzandone $n-1$ per la costruzione di un modello 'parziale' per la classificazione degli utenti; tale modello sarà di conseguenza differente

rispetto al modello da utilizzare nella fase di test ricavato in precedenza, essendo stato ricavato da un sottoinsieme dell'insieme di apprendimento iniziale. Per ognuno dei modelli parziali viene realizzato un test esaustivo, applicando il modello alla restante partizione dell'insieme di apprendimento iniziale per ciascuna delle possibili combinazioni di parametri e calcolando le metriche di valutazione delle prestazioni ottenute per tutte le combinazioni sperimentate.

Seguendo questo procedimento si ottengono, per ciascuna combinazione di parametri e per ciascuna delle metriche di valutazione, n valori dei quali viene calcolata la media in modo da ottenere degli indicatori complessivi delle performance ottenute, i quali saranno utilizzati per la scelta della migliore combinazione di parametri.

La Figura 4.3 mostra un esempio dell'applicazione della n-fold Cross Validation con n uguale a 3; in questo caso si utilizzano due terzi dell'insieme di apprendimento iniziale per ricavare i modelli parziali, mentre il restante terzo viene utilizzato per effettuare i test esaustivi e ricavare gli indicatori delle performance ottenute per ciascuna combinazione di parametri.

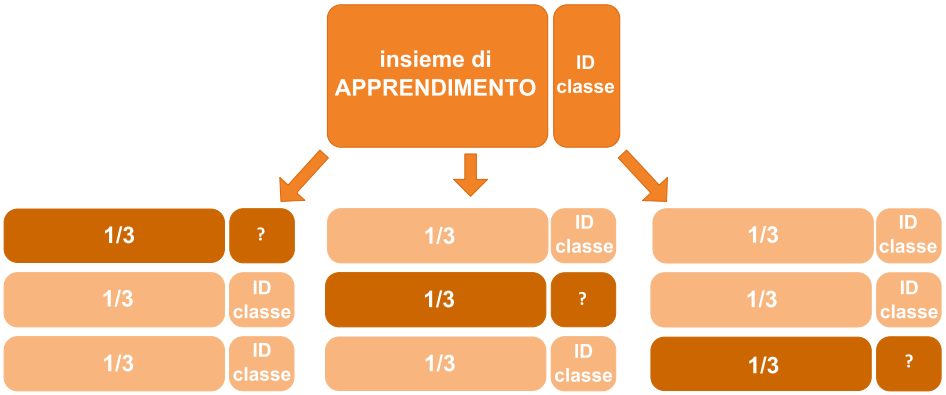


Figura 4.3: *un esempio di 3-fold Cross Validation.*

La fase successiva dello schema implementativo prevede l'applicazione del modello e della corrispondente migliore combinazione di parametri agli utenti dell'insieme di test, allo scopo di assegnare ad ognuno di essi una classe.
Dal confronto tra la classe assegnata tramite l'applicazione del modello e l'effettiva classe di appartenenza dell'utente si ricavano le statistiche per la valutazione della qualità dell'algoritmo utilizzato.

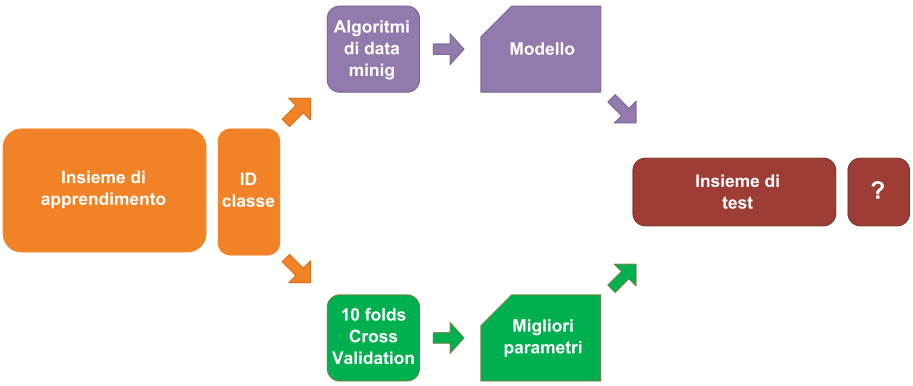


Figura 4.4: *lo schema implementativo applicato per la categorizzazione degli utenti di test.*

Capitolo 5

Algoritmi di base

Indice

In questo capitolo sono presentate le soluzioni ottenute applicando lo schema implementativo proposto nel Capitolo 4 agli algoritmi per la classificazione degli utenti basati sulle Regole di Associazione (Paragrafo 5.1), Singular Value Decomposition (Paragrafo 5.2) e Support Vector Machines (Paragrafo 5.3). Per ognuna delle tre metodologie utilizzate sono descritte le operazioni eseguite per adattare lo schema generale alle peculiarità di ciascun algoritmo, oltre agli applicativi utilizzati.

5.1 Regole di Associazione

Nell'applicazione di questa tecnica il modello per la classificazione è costituito da un insieme di Regole di Associazione, mentre i parametri da considerare sono:

- confidenza: per cui sono previsti valori compresi tra un un minimo di $0{,}01$ ad un massimo di $0{,}59$, con intervalli di $0{,}02$;
- supporto: per il quale sono utilizzati valori compresi tra un minimo pari a 5 ed un massimo pari a 245, con intervalli di 30;

Nel seguito vengono descritte dettagliatamente le operazioni eseguite per effettuare la caratterizzazione degli utenti applicando la metodologia basata sulle regole di associazione. Lo stesso schema implementativo è stato applicato sia per il dataset MovieLens, sia per il dataset Yahoo!.

5.1.1 Preparazione del dataset

Prima di applicare il metodo basato sulle AR sono state effettuate alcune operazioni sulla matrice UGM originale, per renderla adatta allo scopo prefissato. In primo luogo è stata applicata la procedura di normalizzazione, ottenendo così la matrice UGMNorm, in cui la distribuzione delle preferenze degli utenti risulta maggiormente bilanciata rispetto alla matrice non normalizzata. In seguito sono stati individuati i generi preferiti di ciascun utente (1, 2 o 3 a seconda del test effettuato) come descritto nel Paragrafo 4.2.2. Sono anche stati eliminati dai dataset eventuali record incompleti, che avrebbero potuto originare incongruenze nei risultati: in particolare sono stati rimossi tutti quelli utenti per i quali non era presente l'anno di nascita nel dataset Yahoo!, o l'età nel dataset MovieLens.

A ciascun utente è stato successivamente assegnato un identificativo corrispondente alla sua effettiva classe di età, da utilizzare per la costruzione del modello nella fase di apprendimento e per la valutazione delle stime nella fase di individuazione dei migliori parametri e calcolo delle statistiche di valutazione complessiva del test. Concatenando la matrice UGMNorm con il vettore colonna contenente gli identificativi delle classi di appartenenza di ciascun utente è stato ottenuto il dataset su cui applicare la metodologia basata sulle regole di associazione.

Il dataset così ottenuto è stato infine suddiviso nei sottoinsiemi TRAINING SET, costituito dal 70% dei record presenti nel dataset, e TESTING SET, comprendente il restante 30% dei record, da utilizzare rispettivamente nelle fasi di apprendimento e test. In precedenza era stato effettuato un rimescolamento casuale dei record, con lo scopo di bilanciare la presenza delle rilevazioni

meno recenti nei tre sottoinsiemi: infatti i record sono inseriti nelle matrici originali secondo un ordine cronologico e il mancato rimescolamento avrebbe portato ad inserire tutte le rilevazioni meno recenti in un unico sottoinsieme, con possibili conseguenze sui risultati finali.

5.1.2 Costruzione del modello

Come già specificato in precedenza, per questo tipo di analisi il modello è costituito da un insieme di regole di associazione; l'algoritmo per la scoperta di tali regole, descritto nel Paragrafo 2.3.1 prevede una fase di individuazione degli itemset frequenti e una fase di ricerca delle regole all'interno degli insiemi appena individuati.
Per eseguire le due fasi previste dall'algoritmo sono stati utilizzati due software già esistenti e ampiamente diffusi in letteratura:

- *Linear time Closed itemset Miner version 2 (LCM2)*[55],basato sull'omonimo algoritmo [43] per l'individuazione degli itemset frequenti nel dataset da analizzare;

- *Rules*[57], utilizzato per la costruzione delle regole di associazione, a partire dagli itemset frequenti individuati in precedenza.

LCM2 riceve in ingresso i record associati agli utenti del TRAINING SET, individua tutti gli itemset presenti e restituisce come output solo quelli con supporto maggiore di un valore di soglia prefissato (5 nel nostro caso).
Un esempio di comando per LCM2 è il seguente:

fim all *fileInputLCM* 5 *fileOutputLCM*

In questo caso LCM2 inserirà nel file di output tutti gli itemset che compaiono almeno 5 volte nel TRAINING SET.

L'output prodotto da LCM2 viene quindi passato come input a Rules il quale, in maniera analoga, produce tutte le regole di associazione derivate dagli itemset frequenti, restituendo solo quelle aventi un valore di confidenza superiore ad una soglia minima prefissata.
Un esempio di comando per Rules è il seguente:

rules *fileOutputLCM* 0,01 *fileOutputRULES*

il quale restituisce tutte le regole che abbiano un valore di confidenza maggiore o uguale a $0{,}01$, prodotte partendo dagli itemset individuati in precedenza e contenuti nel file *fileOutputLCM*, .

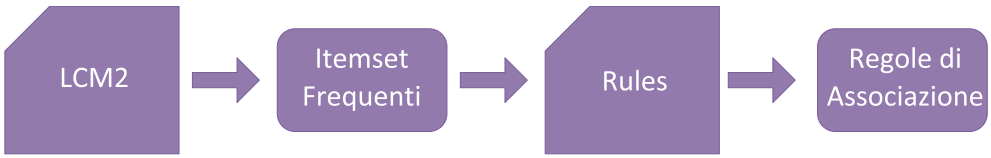


Figura 5.1: *schema di utilizzo di LCM2 e Rules per la creazione del modello basato sulle Regole di Associazione.*

5.1.3 Individuazione dei parametri e test globale

Dopo aver costruito il modello per la categorizzazione degli utenti nel dataset, si è utilizzata la procedura di n-fold Cross Validation per l'individuazione dei parametri migliori. In questo caso è stato utilizzato un valore n pari a 5, a causa di problemi computazionali legati all'utilizzo degli applicativi necessari all'individuazione delle regole di associazione.

L'insieme di apprendimento è stato quindi di volta in volta suddiviso in due partizioni, una, contenente l'80% dei record, su cui ricavare un modello parziale, l'altra, contenente il restante 20% dei record, su cui effettuare un test esaustivo per ciascuna combinazione dei parametri specifici del metodo applicato. Per ogni possibile coppia di valori di confidenza e supporto sono state utilizzate nel procedimento di assegnazione del Lifestyle soltanto quelle regole di associazione in grado di soddisfare contemporaneamente i valori dei due parametri, eliminando dal modello tutte le altre regole.
Le regole di associazione ricavate sono state applicate secondo le seguenti modalità:

- se nel record associato all'utente sono presenti tutti gli item appartenenti all'insieme degli antecedenti della regola che si sta analizzando, si assegna all'utente il Lifestyle associato al conseguente della regola;
- se vi sono due o più regole in "contrasto", cioè che assegnano due Lifestyle differenti allo stesso utente, si privilegia la regola con il valore di confidenza più elevato; in caso di ulteriori contrasti si seleziona la regola con maggiore supporto;

- se nessuna regola può essere applicata, cioè se nessuno degli insiemi antecedenti è interamente contenuto nel record, all'utente associato non viene assegnato alcun Lifestyle.

Dopo aver effettuato la categorizzazione degli utenti per ciascuna coppia confidenza-supporto, sono state calcolate le corrispettive metriche di valutazione, effettuando la media tra i valori calcolati per ognuno dei 5 fold previsti, con l'obiettivo di individuare la combinazione di parametri in grado di fornire i risultati migliori. Tale configurazione di parametri è stata infine applicata per inferire il Lifestyle di un insieme di utenti, permettendo così di calcolare le statistiche complessive descritte nel Paragrafo 4.1.

Schema riassuntivo

Il seguente schema riassume i passi seguiti nella fase di realizzazione dei test per la categorizzazione di utenti applicando la metodologia basata sulle regole di associazione, per i dataset Yahoo! e MovieLens.

1. Preparazione del dataset:
   (a) calcolo della matrice UGM;
   (b) normalizzazione per ottenere UGMNorm;
   (c) estrazione dei generi preferiti (da 1 a 3 a seconda del test effettuato) da UGMNorm;
   (d) eliminazione dei record incompleti e rimescolamento casuale;
   (e) suddivisione dei record in TRAINING SET e TESTING SET.
2. Creazione del modello:
   (a) individuazione degli itemset frequenti nell'insieme di apprendimento con il tool Lcm2;
   (b) individuazione delle regole di associazione con il tool RULES;
3. Individuazione dei parametri tramite il metodo 5-fold Cross Validation:
   (a) costruzione del modello parziale partendo dall'80% dei record contenuti nell'insieme di apprendimento;

(b) per ogni valore di confidenza e supporto assegnazione del Lifestyle al 20% degli utenti, basandosi sul modello parziale ricavato in precedenza;

(c) calcolo delle statistiche di valutazione per ogni coppia di parametri ed individuazione della migliore combinazione.

4. Test globale:

(a) applicazione del modello all'insieme di test, utilizzando la configurazione di parametri migliore;

(b) calcolo delle statistiche complessive per la valutazione del test globale.

5.2 Singular Value Decomposition

Il modello da utilizzare per la classificazione degli utenti in questa metodologia è costituito da una matrice ricavata applicando la Singular Value Decomposition all'insieme di apprendimento.

Lo schema implementativo scelto per l'applicazione di SVD differisce in alcuni aspetti significativi da quello seguito per le Regole di Associazione, descritto nel Paragrafo 5.1. La principale differenza riguarda le modalità con cui è stato descritto il Lifestyle degli utenti nei dataset: mentre nel caso delle Regole di Associazione a ciascun utente veniva associata un'etichetta con l'identificativo della classe di appartenenza, per poter utilizzare la Singular Value Decomposition si è reso necessario rappresentare i Lifestyle degli utenti con una matrice, detta *ActualLifestyle*, la quale è stata in seguito concatenata alla matrice UGMNorm (vedi Paragrafo 2.3.2).

Inoltre, per assegnare gli utenti alle varie classi si è reso necessario l'utilizzo di due matrici: una, detta *EstimatedLifestyle*, contenente i valori ricavati applicando la Singular Value Decomposition all'insieme di test, l'altra, detta *BinaryLifestyle*, contenente il risultato della binarizzazione applicata alla matrice EstimatedLifestyle. Entrambe le matrici appena descritte hanno la stessa struttura di ActualLifestyle.

Di conseguenza, i parametri da considerare nell'applicazione di questo metodo sono;

- *k*: indica il numero di colonne con cui approssimare la matrice modello; sono stati utilizzati valori di *k* compresi tra 1 e 10.

- *t*: indica la soglia di binarizzazione, cioè il valore discriminante per l'assegnazione degli utenti alle classi. I valori utilizzati per questo parametro variano da un minimo di $-0{,}5$ ad un massimo di $0{,}5$, con intervalli pari a $0{,}01$.

5.2.1 Preparazione dei dataset

Le operazioni svolte per la preparazione del dataset sono le medesime utilizzate nell'implementazione del metodo basato sulle Regole di Associazione: dopo aver costruito la matrice UGMNorm ed averne effettuato la normalizzazione per evitare che i risultati fossero influenzati dai generi di film più popolari, si è provveduto all'eliminazione dei generi meno significativi per ciascun utente e dei record incompleti.
Per costruire la matrice ActualLifestyle da concatenare ad UGMNorm, invece, sono state utilizzate le seguenti modalità:

Modalità 1: Inferenza dell'ETÁ

Nei test relativi alla caratterizzazione secondo l'ETÁ si è scelto di esprimere il Lifestyle di ciascun utente utilizzando un vettore riga, con numero di colonne pari al numero di classi di età previste dal test. Tale vettore è stato ottenuto assegnando un valore pari a 0 a tutti gli elementi ad eccezione di quello situato nella posizione corrispondente alla classe di appartenenza dell'utente, cui è stato assegnato un valore pari a 2.

Ad esempio, nel caso in cui fossero previste quattro classi di età, se un utente appartenesse alla classe GIOVANE e a tale classe fosse associata la quarta colonna, il vettore ottenuto sarebbe il seguente:

$$\text{Lifestyle(u)} \quad = \quad \begin{bmatrix} 0 & 0 & 0 & 2 \end{bmatrix}$$

La matrice ActualLifestyle è stata ottenuta concatenando tutti i vettori associati ai singoli utenti, in modo da creare una struttura con numero di righe pari al numero di utenti e numero di colonne pari al numero di classi di età.

Un esempio di matrice per la descrizione del Lifestyle degli utenti in questo tipo di test è il seguente in cui sono previste quattro classi di età:

$$\text{ActualLifestyle} \quad = \quad \begin{bmatrix} 0 & 0 & 2 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 2 \\ & \cdots\cdots & & \\ 0 & 2 & 0 & 0 \end{bmatrix}$$

Modalità 2: Inferenza del GENERE

Per la classificazione degli utenti secondo il GENERE, invece, sono state proposte due alternative per la descrizione del Lifestyle:

1. ogni utente è stato associato ad un'unica etichetta, con valore pari a 2 per gli appartenenti alla classe MASCHIO o a -2 per gli utenti della classe FEMMINA.

2. il Lifestyle di ogni utente è stato rappresentato utilizzando vettori formati da due colonne, analoghi a quelli utilizzati per la categorizzazione secondo classi di età: in questo caso la matrice ActualLifestyle è formata da due colonne, associate rispettivamente alle classi MASCHIO e FEMMINA.

Vengono proposte due possibili matrici ActualLifestyle, per le due modalità appena descritte:

$$\text{ActualLifestyle} = \begin{bmatrix} +2 \\ -2 \\ -2 \\ \cdots \\ -2 \end{bmatrix} \qquad \text{ActualLifestyle} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \\ 0 & 2 \\ & \cdots \\ 0 & 2 \end{bmatrix}$$

Le matrici ottenute dalla concatenazione di UGMNorm con i possibili ActualLifestyle, denominate rispettivamente *UGMNormForGender* e *UGMNormForAge*, hanno costituito la base su cui applicare la Singular Value Decomposition nelle due modalità di test previste.

I record di ciascuna delle due matrici sono infine stati rimescolati casualmente, in modo analogo a quanto svolto per le Regole di Associazione, e successivamente suddivisi negli insiemi TRAINING e TESTING.

5.2.2 Costruzione del modello

La creazione della matrice modello (chiamata Matrix_Model) da utilizzare per la stima del Lifestyle degli utenti del TRAINING SET, è stata ottenuta utilizzando il metodo nativo di MATLAB *svds*:

$$\left[\ U, S, V\ \right] = \text{svds(TRAINING\_SET, n)}$$

il quale esegue la decomposizione sull'insieme di apprendimento, elenca in ordine decrescente i valori singolari individuati e calcola le matrici di output considerando solo i primi n valori singolari di Σ e le prime n colonne di U e V. La matrice V così ottenuta rappresenta il modello da utilizzare nelle fasi successive per la caratterizzazione degli utenti.

$$Matrix\_Model = V$$

5.2.3 Individuazione dei parametri e test globale

L'individuazione dei parametri è stata effettuata applicando la 10-fold Cross Validation; per ciascun fold è stata ricavata una matrice modello, applicando la SVD come descritta nel Paragrafo 2.3.2, che è stata in seguito utilizzata per ottenere la matrice EstimatedLifestyle, applicando tutte le possibili coppie di parametri k e t. A partire da tale matrice è stato possibile ricavare la matrice BinaryLifestyle applicando la regola seguente:
per ogni colonna j della riga rappresentante l'i-esimo utente,

$$\begin{aligned} \text{se} \quad & EstimatedLifestyle_{(i,j)} > t, \\ \text{allora} \quad & BinaryLifestyle_{(i,j)} = 2, \\ \text{altrimenti} \quad & BinaryLifestyle_{(i,j)} = 0. \end{aligned}$$

La caratterizzazione degli utenti è stata eseguita a partire dalla matrice BinaryLifestyle, applicando i seguenti criteri per ciascuna riga:

- se $\forall j\ BinaryLifestyle_{(i,j)} = 0$, non è stata assegnata nessuna classe all'utente i;

- se esiste un unico *j* tale che $BinaryLifestyle_{(i,j)} = 2$, l'utente *i* è stato assegnato alla classe corrispondente alla colonna *j*-esima;

- se esistono due o più *j* tali che $BinaryLifestyle_{(i,j)} = 2$, l'utente *i* è stato assegnato alla classe corrispondente alla colonna $\hat{j}$-esima, tale che $BinaryLifestyle_{(i,\hat{j})}$ sia il valore più vicino alla soglia *t*.

Nella modalità in cui il genere degli utenti è stato rappresentato utilizzando un'unica colonna si è applicata una regola leggermente differente:

$$\begin{gathered}\text{se } EstimatedLifestyle_{(j)} > t,\\ \text{allora } BinaryLifestyle_{(j)} = 2,\\ \text{altrimenti } BinaryLifestyle_{(j)} = -2.\end{gathered}$$

Assegnando di conseguenza alla classe MASCHIO, tutti gli utenti per i quali $BinaryLifestyle_{(j)} = 2$ e alla classe FEMMINA tutti gli utenti per i quali $BinaryLifestyle_{(j)} = -2$.

In corrispondenza di ciascuna coppia (*k*,*t*) sono state quindi ottenute le statistiche di valutazione, calcolando la media dei risultati relativi a ciascun fold, con l'obiettivo di invidiuare la combinazione di parametri che fornisse i risultati migliori in termini di Recall o Lift.

Il modello complessivo ricavato è stato utilizzato per categorizzare gli utenti dell'insieme di test, approssimando la matrice modello con il valore k_{best}, moltiplicando i vettori associati agli utenti per ottenere il Lifestyle ed effettuando la binarizzazione in base al valore t_{best}. I Lifestyle così assegnati sono stati confrontati con le reali classi di appartenenza degli utenti, in modo da calcolare le medesime statistiche di valutazione previste per le Regole di Associazione. La figura 5.2 mostra il processo di assegnazione del Lifestyle nel caso di test sull'età degli utenti; un procedimento analogo è utilizzato per l'assegnazione dei Lifestyle nel caso di test basati sul genere.

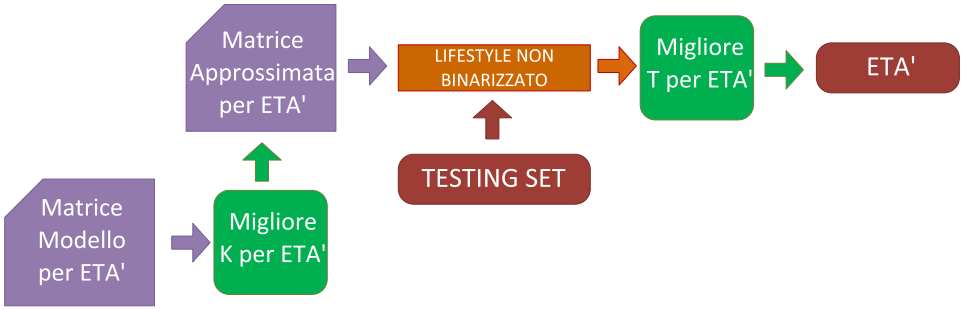


Figura 5.2: *il processo di assegnazione del Lifestyle agli utenti appartenenti a TESTING SET, utilizzando il modello ricavato nella fase precedente.*

Schema riassuntivo

Il seguente schema riassume i passi seguiti nella fase di realizzazione dei test per la categorizzazione di utenti applicando la metodologia basata sulla Singular Value Decomposition, per i dataset Yahoo! e MovieLens.

1. Preparazione del dataset:
   (a) calcolo della matrice UGM
   (b) normalizzazione per ottenere UGMNorm;
   (c) estrazione dei generi preferiti (da 1 a 3 a seconda del test effettuato) da UGMNorm;
   (d) eliminazione dei record incompleti e rimescolamento casuale;
   (e) suddivisione dei record in TRAINING SET e TESTING SET.
2. Creazione del modello:
   (a) calcolo della matrice modello con il metodo svds.
3. Individuazione dei parametri tramite il metodo 10-fold Cross Validation:
   (a) costruzione del modello parziale partendo dal 90% dei record contenuti nell'insieme di apprendimento;
   (b) per ogni valore di k e t assegnazione del Lifestyle al 10% degli utenti dell'insieme di apprendimento, basandosi sul modello parziale ricavato in precedenza;

(c) calcolo delle statistiche di valutazione per ogni coppia di parametri ed individuazione della migliore combinazione.

4. Test globale:

   (a) applicazione del modello all'insieme di test, utilizzando la configurazione di parametri migliore;

   (b) calcolo delle statistiche complessive per la valutazione del test globale.

5.3 Support Vector Machines

L'interazione con una SVM avviene passando come input una serie di valori numerici, nel nostro caso la somma dei rating che ciascun utente ha assegnato ad una determinata categoria, riportata nella matrice UGM.

Il training viene eseguito scegliendo di affidarsi ad un particolare kernel al quale sarà legata la selezione di opportuni parametri. Si effettuerà, attraverso il processo di Cross Validation, la scelta del solo parametro C nel caso si utilizzi un kernel di tipo lineare, di entrambi i parametri C e γ qualora si decida invece di far riferimento ad un kernel di tipo radiale (altri tipi di kernel non saranno trattati all'interno di questa tesi); in entrambe le situazioni la scelta ricadrà sui parametri che garantiscono la miglior Recall globale.

La fase di training porterà alla creazione di un modello che sarà relativo all'iperpiano di separazione migliore e che verrà adoperato nella fase di test per la predizione dei nuovi esempi.

Sommarizzando, una buona procedura di utilizzo di una SVM si articola quindi nei seguenti passi (lo stesso schema è stato applicato sia per il dataset Yahoo! che per il dataset MovieLens):

- trasformare i dati in formato input per SVM (nel caso in cui siano espressi sotto forma di stringhe, è necessario convertirli in formato numerico);
- effettuare uno Scaling dei dati (*opzionale);
- scegliere il tipo di kernel da utilizzare nella fase di training;
- utilizzare la cross validation per la scelta del miglior parametro C ed, eventualmente, γ;

- utilizzare i migliori parametri C e γ ottenuti al punto precedente per 'istruire' il classificatore;
- effettuare il test.

Effettuare uno *Scaling* dei dati, significa passare da un certo intervallo di valori più ampio ad uno più ristretto, operazione questa che consente sia di gestire meglio i calcoli matematici (ottenendo quindi una maggiore accuratezza), sia di evitare attribuiti con un range numerico troppo vasto. Ad esempio, è possibile scalare l'intervallo $[-10, 10]$ in uno più ridotto di tipo $[-1, 1]$. L'utilizzo dello Scaling rimane comunque opzionale (ma consigliabile), in quanto una SVM è in grado di operare con dati espressi in qualsiasi tipo di intervallo.

5.3.1 Preparazione del Dataset

Come nel caso dei due algoritmi precedenti, possono essere effettuate delle operazioni sulla matrice UGM originale prima di passare alla fase di applicazione dell'algoritmo vera e propria. Questi accorgimenti, che nel caso delle Regole di Associazione e di SVD venivano eseguiti di norma, con SVM sono da ritenersi opzionali, perchè, come si vedrà poi nella parte relativa ai risultati, non sempre la loro esecuzione garantisce un incremento positivo delle performance.

In primo luogo, è possibile applicare la procedura di normalizzazione, ottenendo la già citata matrice UGMnorm, in modo tale da bilanciare in maniera più accurata le preferenze espresse dagli utenti per i vari generi dei film rispetto alla matrice originale.

Si può dunque procedere con l'eliminazione dei valori associati ai generi ritenuti meno significativi dall'utente, così da ottenere una versione della UGMnorm che faccia riferimento soltanto all'ambito o agli ambiti (come sempre, è possibile specificare un valore a riguardo che varia da 1 a 3) preferiti.

Se le due operazioni di normalizzazione e di specifica delle preferenze sono opzionali nel caso di SVM, non vale la stessa cosa per quanto riguarda l'eliminazione di eventuali record incompleti, come quelli per cui non è presente l'anno di nascita nel dataset Yahoo! o l'età nel dataset MovieLens; questa operazione deve infatti essere eseguita prima dell'applicazione dell'algoritmo, per evitare l'insorgere di errori nei risultati.

Il dataset risultante viene quindi diviso nei sottoinsiemi TRAINING SET (70%) e TESTING SET (30%), da utilizzare rispettivamente nella fase di apprendimento e di valutatazione delle stime; la divisione è preceduta da un rimescolamento casuale dei record tale da bilanciare la presenza delle rilevazioni meno recenti nei sottoinsiemi ottenuti, come spiegato precedentemente.

5.3.2 Costruzione del modello

La porzione di TRAINING SET ottenuta è destinata a fornire un campo di apprendimento per la SVM, la quale può però essere predisposta in diversi modi durante la fase di learning. La scelta dei migliori parametri di configurazione viene affidata al procedimento di Cross Validation, all'inizio del quale il TRAINING SET viene ulteriormente diviso in 10 fold; ognuno di essi interagisce a turno con una SVM invocata di volta in volta dall'intera procedura.

Si può scegliere di eseguire l'operazione scalando i dati in un intervallo ben preciso (durante il nostro studio abbiamo utilizzato il range $(-1, 1)$ oppure $(0, 1)$), in modo tale da 'avvicinare' tra loro i vari ratings e rendere le operazioni matematiche più accurate e meno dispendiose in termini di calcolo; per ogni elemento viene messo a punto infatti il seguente adattamento:

$$V_{new} = l_B + (u_B - l_B) \cdot ((V_{old} - min)/(max - min))$$

dove V_{new} e V_{old} sono rispettivamente il valore numerico nel set prima e dopo la procedura di Scaling, l_B e u_B i limiti inferiore e superiore del range scelto, mentre min e max indicano il minimo e il massimo valore presente all'interno del sottoinsieme che si sta considerando.

Al termine dell'intero procedimento di Cross Validation si può disporre dei parametri che hanno garantito la miglior Recall globale su tutti e 10 i fold e con i quali è possibile predisporre accuratamente la macchina per l'apprendimento relativo al TRAINING SET.

Bisogna specificare che una SVM prende dati in ingresso secondo uno specifico formato standard: nei vari sottoinsiemi ricavati dalla UGMnorm originale, i ratings sono già espressi in forma numerica come richiesto, ma è necessario che la relazione tra di essi e le varie categorie di film, nonchè la label associata a ciascun training point del dataset, sia riportata in una maniera ben precisa.

I sottoinsiemi della matrice UGMnorm e le label corrispondenti vengono dunque passati in input ad un parsificatore che produce stringhe relative a ciascun utente di tipo:

$$\begin{array}{l} \mathbf{1}\ \ 1:0.005562\ \ 2:0.002545\ \ 3:0.006633\ \ 4:0.002628..... \\ \mathbf{-1}\ \ 1:0.000506\ \ 2:0.000364\ \ 3:0.002673\ \ 4:0.000219..... \end{array} \tag{5.1}$$

Come è possibile notare dalla (5.1), il parsificatore pone all'inizio della stringa l'etichetta della classe a cui l'utente appartiene (evidenziata in grassetto), seguita dall'identificativo numerico di ogni categoria di film (che la SVM identifica come *feature*) e dal rating globale ad essa associato, che segue immediatamente il carattere separatore ':'.

Dopo aver passato in input i dati al parsificatore, le informazioni dell'insieme di apprendimento e di quello di valutazione assumono dunque la forma espressa dalla (5.1).

Entrambi i set devono essere analizzati e valutati da un apposito tool che implementa il comportamento di una specifica macchina SVM. All'inizio del nostro percorso di tesi abbiamo preso in considerazione sia l'utilizzo del pacchetto fornito da *LibSvm*, sia quello proposto da *SVM-light*, scegliendo in seguito di affidarci alle funzionalità messe a disposizione dal primo, dimostratosi fin da subito più rapido e flessibile, oltre ad offrire un numero superiore di servizi. LibSvm dispone di una serie di applicativi, tra cui *Svm-train* relativo alla fase di learning e di elaborazione di un modello e *Svm-predict* per la fase di testing.

Svm-train esamina le varie stringhe parsate del TRAINING SET sulla base delle specifiche richieste. Oltre alla selezione del tipo di kernel, è infatti possibile scegliere di operare in modalità *C-classification* o *Nu-Classification*.

La *C*-classification agisce seguendo l'impostazione teorica classica di una SVM, come illustrato precedentemente (C è infatti il classico parametro relativo alla penalità d'errore), mentre la *Nu*-classification introduce un nuovo parametro ν, il cui valore varia nell'intervallo $(0, 1]$, che va a sostituire C e rappresenta un lower ed un upper bound sul numero di esempi che sono support vector ma giacciono dalla parte sbagliata dell'iperpiano.

Il limite inferiore di tale range va regolato specificatamente ad ogni problema binario secondo la relazione:

$$\nu \leq \frac{2 \cdot min(\#y_i = +1, \#y_i = -1)}{l} \leq 1$$

In questa modalità, considerati i training vector $x_i \in R^n,\ i = 1, ..., l$ appartenenti a due differenti classi e un vettore $\mathbf{y} \in R^l$ tale che $y_i \in \{-1, 1\}$, il problema di ottimizzazione si traduce nella forma

$$\frac{1}{2}\mathbf{w}^T\mathbf{w} - \nu\rho + \frac{1}{l}\sum_{i=1}^{l}\xi_i \tag{5.2}$$

soggetta ai vincoli

$$y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq \rho - \xi_i,$$

$$\xi_i \geq 0,\ i = 1, ..., l,\ \rho \geq 0$$

dove, oltre al parametro ν, compare una nuova variabile ρ che deve essere ottimizzata, mentre $\phi(\mathbf{x}_i)$ mappa x_i in uno spazio dimensionalmente superiore, come già riportato precendemente.

Per capire il ruolo svolto da ρ, si può notare che per $\boldsymbol{\xi} = 0$ la (5.2) stabilisce semplicemente che le due classi sono separate dal margine $\frac{2\rho}{||\mathbf{w}||}$.

Svm-train elabora quindi le informazioni in suo possesso sulla base delle impostazioni scelte dall'utente e produce come output un determinato modello composto da una successione di stringhe; le prime riguardano le opzioni selezionate durante la fase di interazione con l'applicativo, mentre le rimanenti rappresentano i support vector riportati secondo l'ordine imposto dalle label. Se si hanno k classi (con $k \geq 2$) ad esempio, i support vector saranno disposti secondo un ordine di etichetta crescente dalla prima alla $k-$esima classe.

I vettori di supporto sono quindi accompagnati (considerando sempre un contesto composto da k classi) da $k - 1$ coefficienti $y_i \cdot \alpha_i$ dove i vari α_i sono le soluzioni duali dei problemi binari

$$1\ vs\ j,\ 2\ vs\ j,\ ...,\ j-1\ vs\ j,\ j\ vs\ j+1,\ j\ vs\ j+2,\ ...,\ j\ vs\ k$$

e dove vale $y = 1$ per i primi $j - 1$ coefficienti, $y = -1$ per i restanti $k - j$ coefficienti. Nel caso in cui k sia uguale a 4, il model file assumerà quindi una struttura simile a quella della figura 5.3.

Figura 5.3: *Schema del modello prodotto dall'applicativo Svm-train in un contesto di 4 classi: i support vector e le informazioni sui coefficienti, soluzioni dei vari problemi binari, sono riportati secondo l'ordine delle classi a cui appartengono i vari training point*

5.3.3 Individuazione dei parametri e test globale

Dopo aver diviso il dataset originario in TRAINING e TESTING SET ed avere parsato i dati, si deve procedere utilizzando LibSvm in modalità $C-$ o $Nu-$Classification e scegliere uno specifico kernel.

E' necessario tenere presente che deve essere adoperata la stessa configurazione *modalità-kernel* anche nella procedura di Cross Validation, durante la quale i parametri C e γ associati al kernel variano rispettivamente nell'intervallo $(2^{-3}, 2^{15})$ e $(2^{-15}, 2^{3})$ sino alla scelta dei valori che garantiscono la migliore Recall globale su tutti i 10 fold. Gli estremi dei due intervalli sono stati scelti in modo tale che valori maggiori del limite superiore o minori di quello inferiore non diano risultati significativamente diversi da quelli prodotti facendo riferimento ad essi.

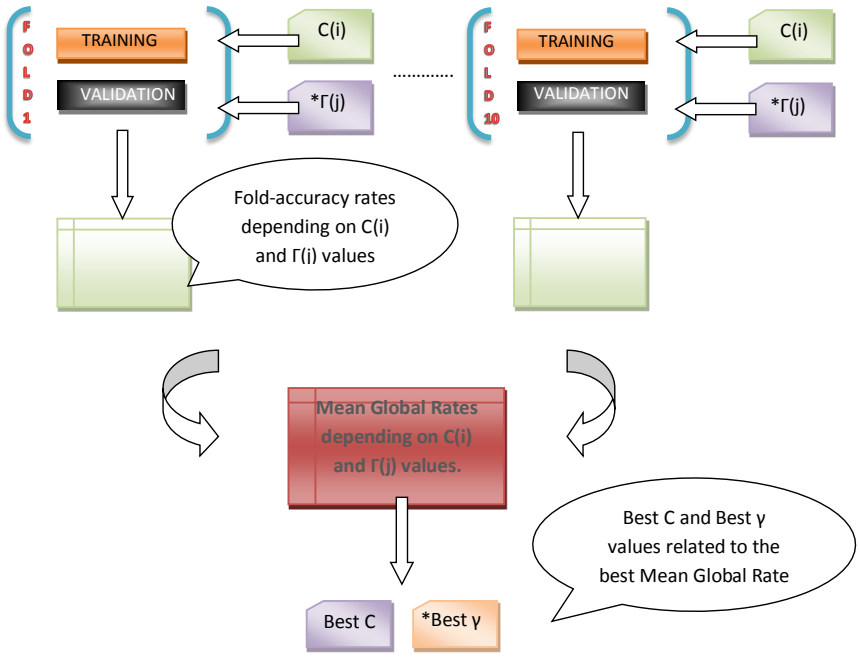


Figura 5.4: *10-folds Cross Validation: il TRAINING SET viene diviso in 10 fold per ognuno dei quali viene effettuata una fase di apprendimento e di test variando i valori di C e γ (solo nel caso di kernel radiale), per stilare i parametri che garantiscono la miglior Recall globale su tutti i fold.*

Dopo la fase di selezione dei migliori parametri, si procede con il learning

relativo all'intero TRAINING SET, invocando Svm-train con la stessa configurazione adoperata durante la Cross validation ed utilizzando i migliori C e γ (in caso di scelta di kernel radiale) ricavati da essa; l'output risultante è un modello dello stesso tipo di quello della figura 5.3.

In base ai parametri contenuti nel modello, è possibile dar luogo all'iperpiano di separazione migliore tramite cui passare al vaglio i vari punti del TESTING SET, in modo da predire la loro classe di appartenenza. In questa fase viene invocato il tool Svm-predict che prende in input la porzione dell'insieme di test parsato ed il modello generato da Svm-train, fornendo come output un vettore con le etichette relative alla classe di appartenenza predetta per ciascun esempio del set.

Il motivo per cui Svm-predict prende in ingresso un formato parsato con tanto di label anteposte alle varie stringhe come espresso dalla (5.1) è che esso calcola e stampa direttamente a video il valore della Accuracy globale dell'insieme di dati analizzato. Poichè il nostro obiettivo consiste nello scendere maggiormente nel dettaglio nell'analizzare i risultati ottenuti, non ci siamo fermati al valore di accuratezza fornito dal tool, ma ci siamo preoccupati di effettuare, al termine della predizione, le consuete valutazioni ricavando, oltre al lift, informazioni sulle Accuracy e Recall delle singole classi, come verrà mostrato nella parte relativa ai risultati.

Nota: in presenza di più classi, lo schema implementativo rimane invariato, dato che LibSvm è in grado di risolvere il problema originario suddividendolo in sottoproblemi binari, basandosi sulle etichette fornite assieme al TRAINING e TESTING SET e sul metodo di Pairwise classification.

Schema riassuntivo

Viene ora proposto uno schema riassuntivo dei vari step di realizzazione dei test per la categorizzazione degli utenti attraverso l'impiego di una SVM, valido sia per il dataset di Yahoo! che per quello di MovieLens.

1. Preparazione del Dataset:
   (a) calcolo della matrice UGM;
   (b) eventuale normalizzazione per ottenere UGMnorm;
   (c) eventuale estrazione dei generi preferiti (da 1 a 3);

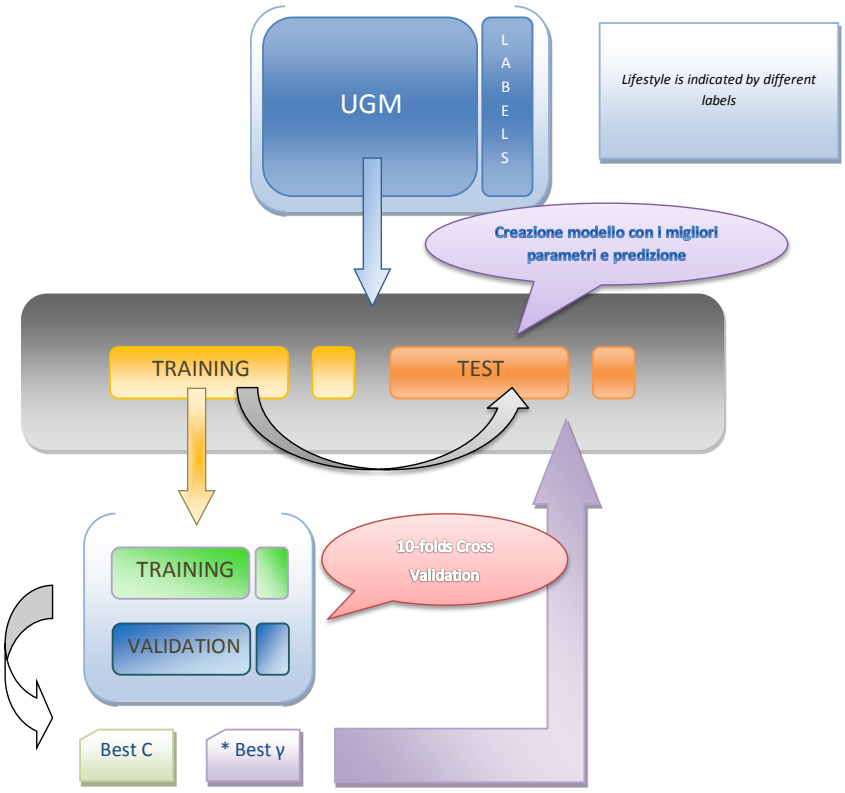


Figura 5.5: *Schema riassuntivo di implementazione per una macchina SVM*

(d) eliminazione dei record incompleti e rimescolamento casuale;

(e) suddivisione dei record in TRAINING SET e TESTING SET.

2. 10-folds Cross Validation, LibSvm e realizzazione del modello:

   (a) scelta del tipo di modalità e di kernel da utilizzare nell'intero processo;

   (b) 10-folds Cross validation con eventuale Scaling dei dati dei vari fold parsati opportunamente all'interno;

   (c) eventuale Scaling del TRAINING e del TESTING SET;

   (d) parsificazione dei dati del TRAINING e del TESTING SET;

   (e) invocazione del tool Svm-train con la configurazione *modalità-kernel* prestabilita ed i parametri ottenuti dalla Cross validation e creazione del modello relativo al TRAINING SET (fase di apprendimento).

3. Predizione e analisi dei risultati:

   (a) invocazione del tool Svm-predict e analisi del TESTING SET sulla base del modello prodotto durante la fase precedente;

   (b) calcolo delle statistiche globali e delle singole classi.

Capitolo 6

Risultati ottenuti applicando gli algoritmi di base

Indice

Questo capitolo contiene i risultati ottenuti nei test sull'inferenza del Lifestyle degli utenti appartenenti ai dataset Yahoo! e MovieLens; per entrambi sono proposti i risultati relativi alle seguenti tipologie di test:

1. inferenza del *sesso* degli utenti, che vengono quindi suddivisi nelle classi MASCHI e FEMMINE, utilizzando i seguenti algoritmi:

- Regole di Associazione
- Singular Value Decomposition
- Support Vector Machines

2. inferenza dell'*età* degli utenti, utilizzando i seguenti algoritmi:

   - Regole di Associazione
   - Singular Value Decomposition
   - Support Vector Machines

In particolare il Paragrafo 6.1 contiene i risultati ottenuti per i test effettuati sul dataset MovieLens, mentre i risultati relativi alla categorizzazione degli utenti del dataset Yahoo! sono presentati nel Paragrafo 6.2.

Per ciascuna delle tipologie di test eseguite sono riportate le configurazioni di parametri che consentono di ottenere le migliori performance in termini di Lift e Recall globale, soffermandosi non solo sui risultati relativi alle metriche generali, ma considerando anche quelli associati alle singole classi previste.

Relativamente ad AR e SVD sono stati effettuati test considerando un numero limitato di generi (uno,due o tre a seconda del test), scelti individuando i preferiti dell'utente. In questo capitolo sono riportati solamente i risultati relativi ai test in cui si considerano tre generi di film, che si dimostrano essere i migliori tra tutti quelli effettuati. I risultati dei test relativi ad uno e due generi di film preferiti sono mostrati nell'Appendice A.

Il Paragrafo 6.3, infine, contiene un confronto tra i migliori risultati ottenuti per l'inferenza del sesso e dell'età utilizzando i tre algoritmi proposti; anche in questo caso sono riportati i risultati relativi ad entrambi i dataset analizzati.

6.1 Dataset MovieLens

In questo paragrafo sono presentati i risultati più significativi dei test svolti per la caratterizzazione degli utenti appartenenti al dataset MovieLens. Sono state eseguite due differenti tipologie di test, basate rispettivamente sull'inferenza del sesso e dell'età degli individui.

6.1.1 Risultati relativi all'inferenza del sesso

Come già evidenziato nel Paragrafo 3.1, il dataset MovieLens contiene informazioni demografiche relative a 4 331 utenti maschi, corrispondenti al $71{,}7\%$ del totale, e a 1 709 utenti femmine, che costituiscono il restante $28,3\%$.

Regole Di Associazione

I migliori risultati, sia in termini di Lift che in termini di Recall Globale, si ottengono applicando valori di confidenza e supporto pari rispettivamente a $0{,}51$ e 35 e sono riportati nella Tabella 6.1.

	METRICA	MASCHI [0,7170]	FEMMINE [0,2830]	TOTALE
Best Lift e Recall globale	Recall	0,9341	0,2790	-
	Accuracy	0,7665	0,6265	-
	Utenti Assegnati	1759	253	2012
	Utenti Non Assegnati	0	0	0
	Regole Generate	94	11	105
	Recall Globale = 0,7488 confidenza = 0,51		Lift = 1,0440 supporto = 35	

Tabella 6.1: *dataset MovieLens - Risultati per la caratterizzazione degli utenti in base al sesso, applicando il metodo basato sulle AR. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

Analizzando i risultati è possibile notare come il Lift ottenuto sia di $1{,}0440$, mentre la Recall complessiva sia pari a $0{,}7488$. Per quanto riguarda le singole classi di utenti, la Recall e la Accuracy relative alla classe MASCHI sono pari a $0{,}9341$ e $0{,}7665$, entrambi superiori alla percentuale di utenti che la compongono. Relativamente alla classe FEMMINE, invece, la Recall ottenuta è leggermente inferiore alla percentuale di utenti appartenenti alla classe (pari a circa il 28%), mentre il valore di Accuracy calcolato è pari a $0{,}6265$, più del doppio rispetto a tale percentuale.

Singular Value Decomposition

I migliori risultati relativi all'inferenza del sesso ottenuti basandosi su questa metodologia si ottengono per la configurazione descritta nella Tabella 6.2, che prevede l'utilizzo di una sola colonna per la descrizione del Lifestyle

degli utenti, ottenendo un Lift pari a $1{,}0484$ ed una Recall globale di $0{,}7533$. Per entrambe le classi si osservano valori di Recall ed Accuracy superiori rispetto alla percentuale di utenti presenti; per la classe MASCHI tali metriche assumono valori pari a $0{,}8617$ e $0{,}8078$, rispettivamente. Per quanto riguarda la classe FEMMINE, invece, si ottengono valori pari a $0{,}4765$ per la Recall e $0{,}5745$ per la Accuracy. Il valore di Accuracy complessivo è lo stesso registrato per la Recall, avendo effettuato una previsione sulla classe di appartenenza per tutti gli utenti dell'insieme di test.

	METRICA	MASCHI [0,7170]	FEMMINE [0,2830]	TOTALE -
Best Lift e Recall Globale	Recall	0,8617	0,4765	-
	Accuracy	0,8078	0,5745	-
	Utenti Assegnati	1389	423	1812
	Utenti Non Assegnati	0	0	0
	Recall Globale = 0,7533		Lift = 1,0484	
	$k_{best} = 7$		$t_{best} = -0{,}12$	

Tabella 6.2: *dataset MovieLens - Risultati per la caratterizzazione degli utenti in base al sesso, applicando il metodo basato sulla SVD. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

La migliore combinazione di parametri da applicare nel caso dell'utilizzo di due colonne, invece, da origine ad un Lift pari a $1{,}0261$ e ad una Recall complessiva di $0{,}7373$ ed è ottenuta considerando un valore di *k* pari a 7 ed una soglia *t* uguale a $-0{,}5$, ma non viene mostrata in quanto origina risultati inferiori rispetto alla configurazione che utilizza una sola colonna.

Support Vector Machines

Come già anticipato nel capitolo 5, il tool LibSvm adoperato può essere impostato secondo diverse configurazioni. Analizziamo nel dettaglio i risultati ottenuti per la modalità *C*-classification espressi nella Tabella 6.3

Come si può notare, in questa modalità i valori più alti in termini di Lift e di Recall globale si ottengono facendo affidamento ad un kernel di tipo lineare, che, come già detto in precedenza, lavora meglio con dataset le cui righe (che nel nostro caso rappresentano gli utenti) sono di gran lunga superiori rispetto al numero di colonne (rappresentanti i generi dei film). Il kernel radiale garantisce comunque buoni risultati, soprattuto nei casi in cui

Matrice Utilizzata	Kernel	Scaling	C	γ	Lift	Recall globale	Accuracy globale
UGMnorm	Lineare	-	32768	-	1,0607	0,7621	0,7621
	Lineare	V [0, 1]	8192	-	1,0538	0,7572	0,7572
	Lineare	A[0, 1]	512	-	1,0736	0,7726	0,7726
	Lineare	V[−1, 1]	32768	-	1,0606	0,7627	0,7627
	Lineare	A[−1, 1]	32768	-	1,0674	0,7693	0,7693
	Radiale	-	32768	8	1,0660	0,7671	0,7671
	Radiale	V[0, 1]	32	8	1,0223	0,7351	0,7351
	Radiale	A[0, 1]	32	8	1,0656	0,7616	0,7616
	Radiale	V[−1, 1]	2048	0.1250	1,0216	0,7301	0,7301
	Radiale	A[−1, 1]	2	8	1,0597	0,7638	0,7638
UGMnoNorm	Lineare	A[0, 1]	8192	-	1,0699	0,7688	0,7688
	Radiale	A[0, 1]	128	8	1,0681	0,7699	0,7699
	Lineare	A[−1, 1]	32768	-	1,0790	0,7765	0,7765
	Radiale	A[−1, 1]	128	8	1,0683	0,7688	0,7688

Tabella 6.3: *Inferenza del sesso sul dataset MovieLens attraverso l'utilizzo di una SVM in modalità C-classification. Le lettere 'V' e 'A' indicano rispettivamente l'utilizzo dello scaling solo durante la procedura di Cross Validation e il caso in cui esso viene applicato anche nella fase di apprendimento e test.*

l'UGM non venga normalizzata, poichè viene reso più efficace l'effetto della procedura di scaling negli specifici intervalli, 'avvicinando' tra di loro i ratings espressi dagli utenti.

Il miglior risultato in termini di Recall globale ($0{,}7765$) si ottiene facendo riferimento ad un kernel di tipo lineare ed applicando uno scaling totale nell'intervallo $(-1, 1)$, con C che assume il valore 32 768; in questo caso le Accuracy e le Recall delle singole classi assumono i valori espressi dalla Tabella 6.4. Tale configurazione permette anche di ottenere il miglior risultato in termini di Lift ($1{,}0790$) per la modalità *C*-classification.

Basandoci sulla modalità *Nu*-classification, invece, sono stati ottenuti i risultati espressi dalla Tabella 6.5. Questa modalità produce le migliori performance in termini di Recall e di Lift, ma anche differenze più marcate tra le varie configurazioni rispetto alla *C*-classification, a seconda della scelta del tipo di kernel e degli intervalli di scaling.

Il miglior risultato in termini di Recall globale ($0{,}7925$), che è anche il migliore in assoluto tra tutti i test di questo tipo eseguiti sul dataset MovieLens, si ottiene facendo riferimento ad un kernel di tipo lineare ed applicando uno scaling totale nell'intervallo $(-1, 1)$, con C uguale a $0{,}125$ e ν pari a $0{,}5$.

	METRICA	MASCHI [0,7170]	FEMMINE [0,2830]
Best Lift e Recall globale	Accuracy	0,7869	0,7102
	Recall	0,9456	0,3425
	Recall globale = 0,7765		Lift = 1,0790
	C = 32768 γ = - UGM normalizzata Kernel: LINEARE Scaling: $[-1, 1]$		

Tabella 6.4: *dataset MovieLens - Risultati per la caratterizzazione degli utenti in base al sesso, applicando il metodo basato su SVM in modalità C-classification. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

Matrice Utilizzata	Kernel	Scaling	C	γ	ν	Lift	Recall globale	Accuracy globale
UGMnorm	Lineare	A[0, 1]	0,125	-	0,50	1,0971	0,7853	0,7853
	Lineare	V[0, 1]	0,125	-	0,50	1,0713	0,7793	0,7793
	Lineare	A[−1, 1]	0,125	-	0,50	1,0821	0,7925	0,7925
	Lineare	V[−1, 1]	0,125	-	0,50	1,0556	0,7544	0,7544
	Radiale	A[0, 1]	0,125	2^{-11}	0,46	1,0299	0,7412	0,7412
	Radiale	V[0, 1]	0,125	2^{-11}	0,39	0,7112	0,5110	0,5110
	Radiale	A[−1, 1]	0,125	2^{-11}	0,44	0,8687	0,6242	0,6242
	Radiale	V[−1, 1]	0,125	0,313	0,49	0,9571	0,6887	0,6887
UGMnoNorm	Radiale	A[0, 1]	0,125	8	0,46	1,0548	0,7644	0,7644
	Lineare	A[0, 1]	0,125	-	0,50	1,0698	0,7699	0,7699
	Radiale	A[−1, 1]	0,125	2	0,47	1,0706	0,7704	0,7704
	Lineare	A[−1, 1]	0,125	-	0,50	1,0869	0,7803	0,7803

Tabella 6.5: *Inferenza del sesso sul dataset MovieLens attraverso l'utilizzo di una SVM in modalità Nu-classification. Le lettere 'V' e 'A' indicano rispettivamente l'utilizzo dello scaling solo durante la procedura di Cross Validation e il caso in cui esso viene applicato anche nella fase di apprendimento e test.*

Anche per quanto riguarda il Lift, il miglior risultato ottenuto per questa configurazione è il migliore tra tutti quelli relativi all'inferenza del sesso eseguiti sul dataset di MovieLens, ed è pari a $(1{,}0971)$. Tale risultato si ottiene facendo riferimento ad un kernel di tipo lineare ed applicando uno scaling totale nell'intervallo $(0, 1)$, con C che assume il valore $0{,}125$ e ν pari a $0{,}50$. La Tabella 6.6 mostra i risultati ottenuti applicando le due configurazioni appena descritte.

	METRICA	MASCHI [0,7170]	FEMMINE [0,2830]
Best Lift	Accuracy	0,7948	0,7316
	Recall	0,9437	0,3864
	Recall globale = 0,7853		Lift = 1,0971
	C = 0,125 ν = 0,5 UGM normalizzata Kernel: LINEARE Scaling: [0, 1]		
Best Recall	Accuracy	0,8082	0,7026
	Recall	0,9397	0,3897
	Recall globale = 0,7925		Lift = 1,0821
	C = 0,125 ν = 0,5 UGM normalizzata Kernel: LINEARE Scaling: [−1, 1]		

Tabella 6.6: *dataset MovieLens - Risultati per la caratterizzazione degli utenti in base al sesso, applicando il metodo basato su SVM in modalità Nu-classification.*

6.1.2 Risultati relativi all'inferenza dell'età

Per il dataset MovieLens viene proposta una suddivisione degli utenti in quattro classi:

- la classe GIOVANISSIMI, che comprende tutti gli utenti di età inferiore a 18 anni, è composta da 225 utenti, pari al 3,68% del totale;
- gli utenti di età compresa tra 18 e 34 anni formano la classe GIOVANI, composta da 3197 utenti, che rappresentano il 52,94% del totale;
- la classe ADULTI, a cui appartengono 2240 utenti, pari al 37,10% del totale, è costituita da tutti gli utenti di età compresa tra 35 e 54 anni;
- il restante 6,28% degli utenti totali forma la classe ANZIANI, comprendente 378 utenti .

Regole di Associazione

In nessuna delle configurazioni di parametri utilizzate sono stati assegnati utenti alle classi GIOVANISSIMI ed ANZIANI, a causa del fatto che, per tali classi, il numero di utenti presenti nel dataset non è sufficiente per creare regole con la confidenza e il supporto minimi richiesti.

Il miglior valore di Lift viene ottenuto utilizzando una confidenza ed un supporto pari rispettivamente a 0,59 e 35; la configurazione che consente di ottenere la migliore Recall globale, invece, è ottenuta considerando valori di confidenza e supporto pari a 0,43 e 35. I risultati complessivi ottenuti per tali combinazioni di parametri sono riportati nella Tabella 6.7.

Nella configurazione che garantisce il migliore Lift circa un quarto degli utenti non viene assegnato ad alcuna classe: di conseguenza il valore della Recall globale è molto inferiore rispetto al migliore ottenuto per questa tipologia di test, essendo pari a 0,4415. I valori di tali metrica relativi alle singole classi sono molto differenti: mentre per la classe GIOVANI si registra un valore pari a 0,7865, di gran lunga superiore rispetto alla percentuale di utenti che compongono la classe, per quanto riguarda la classe ADULTI si ottiene un valore di 0,0689, che non può esserer considerato soddisfacente. La spiegazione di tale risultato risiede nel numero di regole prodotte che, nel caso della classe ADULTI sono solamente due, contro le 32 individuate per la classe GIOVANI. Nonostante questa differenza nel numero di regole

	METRICA	Anziani [0,0628]	Adulti [0,3710]	Giovani [0,5294]	Giovanissimi [0,0368]	TOTALE
Best Lift	Recall	-	0,0689	0,7865	-	-
	Accuracy	-	0,5946	0,5901	-	-
	Utenti Assegnati	0	88	1425	0	1513
	Utenti Non Ass.	49	230	207	13	499
	Regole Generate	0	2	232	0	234
	Recall Globale = 0,4415 confidenza = 0,59				Lift = 1,0862 supporto = 35	
Best Recall	Recall	-	0,3342	0,8371	-	-
	Accuracy	-	0,5054	0,5873	-	-
	Utenti Assegnati	0	493	1519	0	2012
	Utenti Non Ass.	0	0	0	0	0
	Regole Generate	0	43	88	0	131
	Recall Globale = 0,5672 confidenza = 0,43				Lift = 1,071 supporto = 35	

Tabella 6.7: *dataset MovieLens - Risultati per la caratterizzazione degli utenti in base all'età, applicando il metodo basato sulle AR.*

prodotte, le Accuracy corrispondenti alle due classi sono molto simili, segno che alcune delle regole prodotte per la classe GIOVANI non producono un effettivo miglioramento dei risultati ottenuti.

A proposito della configurazione che assicura la migliore Recall globale, invece, è possibile notare come entrambi i valori di Recall relativi alle due classi in esame siano migliori rispetto alla configurazione precedente, a causa dell'uso di criteri meno restrittivi per l'assegnazione degli utenti alle classi, che si traducono in un numero più elevato di regole prodotte. L'uso di tali criteri ha determinato una lieve diminuzione nei valori di Accuracy relativi alle singole classi, maggiormente evidente per i GIOVANI, che passa da un valore superiore al 59% ad uno di poco superiore al 50%.

Singular Value Decomposition

La combinazione di parametri descritta nella Tabella 6.8, consente di ottenere un Lift pari a $1{,}0515$, assegnando utenti a due classi su quattro (ADULTI e GIOVANI). La Tabella 6.8 mostra anche la configurazione che assicura la migliore Recall globale, ottenuta per valori di k e t pari a 6 e $-0{,}5$, che è pari a $0{,}5419$.

Anche per questa tipologia di test la configurazione che assicura il miglior

	METRICA	Anziani [0,0628]	Adulti [0,3710]	Giovani [0,5294]	Giovanissimi [0,0368]	TOTALE
Best Lift	Recall	-	0,3875	0,6310	-	-
	Accuracy	-	0,4802	0,6177	-	-
	Utenti Assegnati	0	531	994	0	1525
	Utenti Non Ass.	13	111	150	13	287
	Recall Globale = 0,4796 $k_{best} = 8$				Lift = 1,0515 $t_{best} = 0{,}5$	
Best Recall	Recall	-	0,5182	0,6896	-	-
	Accuracy	-	0,4743	0,6139	-	-
	Utenti Assegnati	0	719	1093	0	1812
	Utenti Non Ass.	0	0	0	0	0
	Recall Globale = 0,5419 $k_{best} = 5$				Lift = 1,0092 $t_{best} = -0{,}5$	

Tabella 6.8: *dataset MovieLens - Risultati per la caratterizzazione degli utenti in base all'età, applicando il metodo basato su SVD.*

Lift prevede di non assegnare un significativo numero di utenti ad alcuna classe, con l'obiettivo di minimizzare gli errori di classificazione. Per quanto riguarda la classe GIOVANI, la diminuzione del valore di Accuracy (da $0{,}6177$ a $0{,}6123$) è inferiore rispetto al corrispondente aumento di Recall, che passa da un valore di $0{,}6310$ a uno pari a $0{,}6896$, segno che l'uso di criteri meno restrittivi ha contribuito a migliorare le performance relative alla Recall complessiva. Tale effetto è ancora più significativo a proposito della classe ADULTI, per la quale si registra un sensibile incremento della Recall da un valore di $0{,}3875$ ad uno pari a $0{,}5106$, al quale corrisponde una diminuzione dell'Accuracy pari a circa il 4%.

Support Vector Machines

I risultati per la modalità *C*-classification sono riportati nella Tabella 6.9. Il miglior risultato in termini di Lift ($1{,}1920$) si ottiene facendo riferimento ad un kernel di tipo lineare con scaling totale nell'intervallo $(0, 1)$, senza applicare la normalizzazione alla matrice UGM, con C che assume il valore 32 768; per tale configurazione di parametri, tuttavia, si riescono ad assegnare utenti solamente a due classi sulle quattro previste.

Applicando un kernel di tipo radiale, unitamente ad uno scaling totale nell'intervallo $(0, 1)$, e parametri C e γ che assumono il valore 8, si ottiene un Lift inferiore, pari a ($1{,}1323$), ma si è in grado di assegnare alcuni utenti

Matrice Utilizzata	Kernel	Scaling	C	γ	Lift	Recall globale	Accuracy globale	Classi assegnate
UGMnorm	Lineare	-	32768	-	1,1141	0,5982	0,5982	2 su 4
	Lineare	V [0, 1]	32768	-	1,0998	0,5596	0,5596	2 su 4
	Lineare	A[0, 1]	512	-	1,1730	0,5938	0,5938	2 su 4
	Lineare	V[−1, 1]	32768	-	1,1460	0,5905	0,5905	2 su 4
	Lineare	A[−1, 1]	32768	-	1,1264	0,6049	0,6049	2 su 4
	Radiale	-	32768	8	1,1720	0,5999	0,5999	2 su 4
	Radiale	V[0,1]	2048	0,5	1,0640	0,5596	0,5596	2 su 4
	Radiale	A[0,1]	8	8	1,1323	0,5999	0,5999	3 su 4
	Radiale	V[−1,1]	2048	0,125	1,0298	0,5530	0,5530	2 su 4
	Radiale	A[−1,1]	128	0,5	1,1382	0,5966	0,5966	2 su 4
UGMnoNorm	Lineare	A[0, 1]	32768	-	1,1920	0,6010	0,6010	2 su 4
	Radiale	A[0,1]	32768	2	1,0456	0,5690	0,5690	3 su 4
	Radiale	A[−1,1]	8192	0,5	1,1077	0,5844	0,5844	2 su 4
	Lineare	A[−1, 1]	32768	-	1,1305	0,6071	0,6071	3 su 4

Tabella 6.9: *Inferenza dell'età sul dataset MovieLens attraverso l'utilizzo di una SVM in modalità C-classification. Le lettere 'V' e 'A' indicano rispettivamente l'utilizzo dello scaling solo durante la procedura di Cross Validation e il caso in cui esso viene applicato anche nella fase di apprendimento e test.*

alla classe GIOVANISSIMI. I risultati ottenuti per questa configurazione di parametri sono inseriti nella Tabella 6.10, unitamente alla configurazione che consente di ottenere le migliori performance relativamente alla Recall globale. Il miglior risultato in termini di Recall globale ($0{,}6071$) assegna 3 classi su 4 e si ottiene facendo riferimento ad un kernel di tipo lineare ed applicando uno scaling totale nell'intervallo $(-1, 1)$ senza normalizzare UGM, con C che assume il valore 32 768.

I risultati per la modalità *Nu*-classification sono espressi dalla Tabella 6.11.

A differenza di quanto accade nei test relativi all'inferenza dei generi, la *Nu*-classification produce valori di Recall globale e di Lift inferiori rispetto alla *C*-classification, ma consente di stimare tutte le singole classi, cosa che non è possibile invece facendo riferimento alla prima modalità.

Il miglior risultato in termini di Recall globale ($0{,}53366$) che assegna 4 classi su 4, si ottiene facendo riferimento ad un kernel di tipo radiale ed applicando uno scaling totale nell'intervallo $(-1, 1)$, con C, γ e ν che assumono rispettivamente i valori 32 768, 8 e $0{,}10$. Le Accuracy e le Recall delle singole classi assumono i valori espressi dalla Tabella 6.12. Tale configurazione permette anche di ottenere il miglior risultato in termini di Lift ($0{,}9807$) per la modalità *Nu*-classification.

	METRICA	GIOVANISSIMI [0,0629]	GIOVANI [0,3707]	ADULTI [0,5296]	ANZIANI [0,0368]
Best Lift	Accuracy	1,00000	0,6257	0,5558	–
	Recall	0,02500	0,8469	0,4164	–
	Recall globale = 0,5999 Lift = 1,1323				
	C = 8 γ = 8 UGM normalizzata Kernel: RADIALE Scaling: [0, 1]				
Best Recall	Accuracy	0,2500	0,6211	0,5511	–
	Recall	0,0161	0,8281	0,4330	–
	Recall globale = 0,6071 Lift = 1,1305				
	C = 32768 γ = - UGM non normalizzata Kernel: LINEARE Scaling: [−1, 1]				

Tabella 6.10: *dataset MovieLens - Risultati per la caratterizzazione degli utenti in base all'età, applicando il metodo basato sulle SVM in modalità C-classification.*

Matrice Utilizzata	Kernel	Scaling	C	γ	ν	Lift	Recall globale	Accuracy globale	Classi assegnate
UGMnorm	Lineare	-	0,125	-	0,10	0,3844	0,2064	0,2064	4 su 4
	Lineare	A[0, 1]	0,125	-	0,10	0,6458	0,3422	0,3422	4 su 4
	Lineare	V[0, 1]	0,125	-	0,10	0,7646	0,4034	0,4034	4 su 4
	Lineare	A[−1, 1]	0,125	-	0,10	0,5680	0,3019	0,3019	4 su 4
	Lineare	V[−1, 1]	0,125	-	0,10	0,8937	0,4548	0,4548	4 su 4
	Radiale	-	0,125	8	0,10	0,5906	0,3129	0,3129	4 su 4
	Radiale	A[0, 1]	0,125	8	0,10	0,8602	0,4619	0,4619	4 su 4
	Radiale	A[−1, 1]	0,125	8	0,10	0,9807	0,5337	0,5337	4 su 4
UGMnoNorm	Lineare	A[0, 1]	0,125	-	0,10	0,7873	0,4227	0,4227	4 su 4
	Radiale	A[0, 1]	0,125	8	0,10	0,8541	0,4586	0,4586	4 su 4
	Lineare	A[−1, 1]	0,125	-	0,10	0,7009	0,3764	0,3764	4 su 4
	Radiale	A[−1, 1]	0,125	8	0,10	0,8376	0,4498	0,4498	4 su 4

Tabella 6.11: *Inferenza dell'età sul dataset MovieLens attraverso l'utilizzo di una SVM in modalità Nu-classification. Le lettere 'V' e 'A' indicano rispettivamente l'utilizzo dello scaling solo durante la procedura di Cross Validation e il caso in cui esso viene applicato anche nella fase di apprendimento e test.*

	METRICA	GIOVANISSIMI [0,0629]	GIOVANI [0,3707]	ADULTI [0,5296]	ANZIANI [0,0368]
Best Lift e Recall globale	Accuracy	0,1071	0,6043	0,4678	0,1343
	Recall	0,0556	0,6815	0,4275	0,0818
	Recall globale = 0,5337 Lift = 0,9807				
	C = 32768 γ = 8 ν = 0,10 UGM normalizzata Kernel: RADIALE Scaling: $[-1, 1]$				

Tabella 6.12: *dataset MovieLens: risultati per la caratterizzazione degli utenti in base all'età, applicando il metodo basato sulle SVM Nu-classification. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

Test con minore numero di generi

I risultati dei test effettuati con l'eliminazione dalla UGMnorm delle voci associate ai generi di film ritenuti meno significativi dall'utente sono inferiori rispetto a quelli fin qui illustrati, sia nel caso dell'inferenza del sesso che dell'età. É possibile avere un riscontro di questo fatto considerando le stesse configurazioni che nei test standard hanno permesso di ottenere i risultati migliori, facendo però riferimento ad un numero massimo di generi preferiti variabile tra 1 a 3. Utilizzando ad esempio la modalità *Nu*-classification, uno scaling compreso tra -1 e 1 ed un kernel di tipo lineare con un numero massimo di generi preferiti pari a 3, otteniamo una Recall globale uguale a $0{,}74338$ ed un Lift pari a $1{,}0290$. Questi valori sono decisamente inferiori rispetto a quelli ottenuti utilizzando gli stessi settaggi, ma considerando tutti e quanti i 18 generi. Anche i migliori risultati per l'inferenza dell'età, ricavati considerando solo i generi preferiti non possono essere considerati soddisfacenti.

Questo fatto ci permette di comprendere che una SVM lavora meglio nei casi in cui può disporre di un discreto numero di categorie e di un rilevante numero di dati associati ad esse; decrementando infatti l'insieme dei generi preferiti, i risultati peggiorano. Come è lecito aspettarsi, anche il valore del Lift decresce notevolmente.

6.2 Dataset Yahoo!

Le tipologie di test eseguite sul dataset Yahoo! e gli algoritmi utilizzati sono i medesimi già presentati per il dataset MovieLens; nel seguito sono presentati i risultati ottenuti.

6.2.1 Risultati relativi all'inferenza del sesso

La suddivisione tra utenti maschi e femmine all'interno del dataset Yahoo! è molto simile alla ripartizione mostrata in precedenza, relativa al dataset MovieLens: gli utenti della classe MASCHI, infatti, sono 5 436 e corrispondono al $71{,}3\%$ del totale, mentre il restante $28,7\%$ è costituito da utenti femmine.

Regole Di Associazione

Per quanto riguarda la metodologia basata sulle Regole di Associazione, il migliore risultato ottenuto in termini di Lift viene raggiunto utilizzando valori di confidenza e supporto pari rispettivamente a $0{,}59$ e 185. Come mostrato nella Tabella 6.13, applicando tale combinazione di parametri si ottiene anche la migliore performance in termini di Recall Globale, che è pari a $0{,}7328$.

	METRICA	MASCHI [0,7130]	FEMMINE [0,2870]	TOTALE
Best Lift e Recall globale	Recall	0,9516	0,1880	-
	Accuracy	0,7448	0,6093	-
	Utenti Assegnati	2313	224	2357
	Utenti Non Assegnati	0	0	0
	Regole Generate	113	8	121
	Recall Globale = 0,7328		Lift = 1,0271	
	confidenza = 0,59		supporto = 185	

Tabella 6.13: *dataset Yahoo! - Risultati per la caratterizzazione degli utenti in base al sesso, applicando il metodo basato sulle AR. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

Analizzando i risultati è possibile notare come per entrambe le classi il valore di Accuracy ottenuto sia superiore alla percentuale di utenti presenti; in particolare per la classe FEMMINE si registra una Accuracy pari a $0{,}6093$ rispetto ad una percentuale pari a $28{,}7\%$. Per quanto riguarda la Recall, invece,

si registra un ottimo risultato per la classe MASCHI, pari a $0{,}9516$, mentre per la classe FEMMINE sono stati individuati meno di un quinto degli utenti presenti nell'insieme di test; ciò è probabilmente dovuto all'elevato numero di regole generate per la classe MASCHI, per la quale sono state prodotte ben 113 regole, contro le 8 prodotte per la classe FEMMINE.

Singular Value Decomposition

La configurazione per cui si ottengono i migliori risultati in termini di Lift e di Recall globale è quella in cui si utilizza una sola colonna per la descrizione del sesso degli utenti e in cui si scelgono valori di *k* e *t* pari a 2 e $-0{,}48$. Utilizzando tale configurazione il numero di utenti non assegnati ad alcuna classe è pari a zero e si registrano valori di Accuracy per le due classi pari a $0{,}7612$ per i MASCHI e $0{,}5521$ per le FEMMINE, entrambe superiori alla percentuale di utenti presenti; l'Accuracy globale ottenuta è invece pari a $0{,}7314$. Analizzando i risultati ottenuti in termini di Recall, invece, è possibile notare come si ottengano risultati elevati per la classe MASCHI, per cui si ottiene una Recall pari a $0{,}9136$, mentre per la classe FEMMINE si ottiene un valore pari a $0{,}2697$, inferiore rispetto alla percentuale di utenti nella classe. I risultati relativi alle due classi si riflettono in una Recall globale pari a $0{,}7314$, esattamente lo stesso valore ottenuto per l'Accuracy, non essendovi utenti per i quali non sia stata effettuata la categorizzazione. I risultati ottenuti per questo tipo di configurazione sono riportati nella Tabella 6.14.

	METRICA	MASCHI [0,7130]	FEMMINE [0,2870]	TOTALE
Best Lift e Recall globale	Recall	0,9136	0,2697	-
	Accuracy	0,7612	0,5521	-
	Utenti Assegnati	1976	317	2284
	Utenti Non Assegnati	0	0	0
	Recall Globale = 0,7314		Lift = 1,0201	
	$k_{best} = 2$		$t_{best} = -0{,}48$	

Tabella 6.14: *dataset Yahoo! - Risultati per la caratterizzazione degli utenti in base al sesso, applicando il metodo basato su SVD. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

Per quanto riguarda l'utilizzo di due colonne per descrivere il Lifestyle

degli utenti, la configurazione per la quale si ottengono i risultati migliori prevede l'uso di valori di *t* e *k* pari rispettivamente a $-0{,}5$ ed 1. É possibile notare come, applicando tale configurazione, tutti gli utenti siano assegnati alla classe MASCHI e, di conseguenza, si ottenga un valore pari ad 1 per la Recall relativa a tale classe. Il valore ottenuto per la Recall Globale, invece, è pari a $0{,}7130$ e coincide con la percentuale di utenti che compongono la classe MASCHI. La mancata assegnazione di utenti alla classe FEMMINE fa si che il Lift sia pari ad 1.

Support Vector Machines

I risultati ottenuti per la modalità *C*-classification per il dataset di Yahoo! sono contenuti nella Tabella 6.15.

Matrice Utilizzata	Kernel	Scaling	C	γ	Lift	Recall globale	Accuracy globale
UGMnorm	Lineare	-	32768	-	1,0006	0,7205	0,7205
	Lineare	V [0, 1]	32768	-	0,9988	0,7047	0,7047
	Lineare	A[0, 1]	32768	-	1,0257	0,7329	0,7323
	Lineare	V[-1, 1]	32768	-	1,0030	0,7205	0,7205
	Lineare	A[-1, 1]	32768	-	1,0024	0,7161	0,7161
	Radiale	-	32768	8	1,0018	0,7196	0,7196
	Radiale	V[0,1]	32768	8	1,0086	0,7161	0,7161
	Radiale	A[0,1]	32768	8	1,0387	0,7393	0,7393
	Radiale	V[-1,1]	32768	8	1,0018	0,7196	0,7196
	Radiale	A[-1,1]	32768	8	0,9382	0,6706	0,6706
UGMnoNorm	Lineare	A[0, 1]	512	-	0,9994	0,7152	0,7152
	Lineare	A[-1, 1]	32768	-	1,0154	0,7227	0,7227
	Radiale	A[0,1]	32768	8	1,0219	0,7340	0,7340
	Radiale	A[-1,1]	32768	8	1,0231	0,7349	0,7349

Tabella 6.15: *Inferenza del sesso sul dataset Yahoo! attraverso l'utilizzo di una SVM in modalità C-classification. Le lettere 'V' e 'A' indicano rispettivamente l'utilizzo dello scaling solo durante la procedura di Cross Validation e il caso in cui esso viene applicato anche nella fase di apprendimento e test.*

Il miglior risultato in termini di Recall globale ($0{,}7393$) si ottiene facendo riferimento ad un kernel di tipo radiale ed applicando uno scaling totale nell'intervallo $(0, 1)$, con C che assume il valore 32768, mentre γ è pari a 8. Le Accuracy e le Recall delle singole classi assumono i valori espressi dalla Tabella 6.16. Tale configurazione permette anche di ottenere il miglior risultato in termini di Lift ($1{,}0387$) per la modalità *C*-classification.

	METRICA	MASCHI [0,7130]	FEMMINE [0,2870]
Best Lift e Recall globale	Accuracy	0,7571	0,6121
	Recall	0,9330	0,2610
	Recall globale = 0,7393		Lift = 1,0387
	C = 32768 γ = 8 UGM normalizzata Kernel: RADIALE Scaling: [0, 1]		

Tabella 6.16: *dataset Yahoo! - Risultati per la caratterizzazione degli utenti in base al sesso, applicando il metodo basato su SVM in modalità C-classification. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

Basandoci sulla modalità *Nu*-classification, invece sono stati ottenuti i risultati mostrati nella Tabella 6.17.

Matrice Utilizzata	Kernel	Scaling	C	γ	ν	Lift	Recall globale	Accuracy globale
UGMnorm	Lineare	A[0, 1]	0,125	-	0,30	0,4665	0,332	0,332
	Lineare	V[0, 1]	0,125	-	0,50	1,0081	0,7113	0,7113
	Lineare	A[−1, 1]	0,125	-	0,50	0,3906	0,2813	0,2813
	Lineare	V[−1, 1]	0,125	-	0,50	0,5621	0,4077	0,4077
	Radiale	A[0, 1]	0,125	0,125	0,50	1,0460	0,7222	0,7222
	Radiale	A[−1, 1]	0,125	2^{-11}	0,16	0,3721	0,2673	0,2673
UGMnoNorm	Lineare	A[0, 1]	0,125	-	0,50	1,0302	0,7152	0,7152
	Lineare	A[−1, 1]	0,125	-	0,50	1,0152	0,7301	0,7301
	Radiale	A[0, 1]	0,125	8	0,49	1,0341	0,7297	0,7297
	Radiale	A[−1, 1]	0,125	8	0,50	1,0561	0,7332	0,7332

Tabella 6.17: *Inferenza del sesso sul dataset Yahoo! attraverso l'utilizzo di una SVM in modalità Nu-classification. Le lettere 'V' e 'A' indicano rispettivamente l'utilizzo dello scaling solo durante la procedura di Cross Validation e il caso in cui esso viene applicato anche nella fase di apprendimento e test.*

Il miglior risultato in termini di Recall globale (0,73316) si ottiene facendo riferimento ad un kernel di tipo radiale ed applicando uno scaling totale nell'intervallo $(-1, 1)$, utilizzando la matrice UGM. I parametri C, γ e ν assumono rispettivamente i valori 0,125, 8 e 0,50; le Accuracy e le Recall delle singole classi sono riportate nella Tabella 6.18. Tale configurazione permette anche di ottenere il miglior risultato in termini di Lift (1,0561) per la modalità *Nu*-classification.

	METRICA	MASCHI [0,7130]	FEMMINE [0,2870]
Best Lift e Recall globale	Accuracy	0,7359	0,7070
	Recall	0,9603	0,2174
	Recall globale = 0,7332		Lift = 1,0561
	C = 0,125 γ = 8 ν = 0,50 UGM non normalizzata Kernel: RADIALE Scaling: $[-1, 1]$		

Tabella 6.18: *dataset Yahoo! - Risultati per la caratterizzazione degli utenti in base al sesso, applicando il metodo basato su SVM in modalità Nu-classification. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

6.2.2 Risultati relativi all'inferenza dell'età

Per effettuare l'inferenza dell'età degli utenti sono state effettuate cinque tipologie di test, realizzate modificando di volta in volta il numero di classi in cui suddividere gli utenti e gli anni di nascita utilizzati come estremi per tali classi. Nel seguito vengono mostrati i risultati per la tipologia di test per la quale sono stati ottenuti i migliori risultati in termini di Lift, nella quale gli utenti sono stati suddivisi nelle classi ADULTI e GIOVANI, considerando appartenenti alla classe ADULTI i nati prima dell'anno 1980.

Adottando questo tipo di suddivisione la classe ADULTI risulta formata da 3260 utenti, pari al $53{,}9\%$ del totale; la classe GIOVANI, invece, comprende 4357 utenti, che equivalgono al $46{,}1\%$ del totale.

Regole di Associazione

Il valore più alto registrato per il Lift è pari a $1{,}0506$, ottenuto con valori di confidenza e supporto pari a $0{,}51$ e 35; utilizzando gli stessi parametri si ottiene anche la migliore Recall globale, che è pari a $0{,}5654$.

Per la classe ADULTI viene generato un numero di regole doppio rispetto alla classe GIOVANI; ciò si riflette nei valori delle metriche associate alle due classi. Mentre nel caso della classe GIOVANI sia Accuracy che Recall sono superiori alla percentuale di utenti che compongono la classe, per quanto riguarda la classe ADULTI si ottiene una Recall pari a $0{,}3077$, decisamente inferiore rispetto a tale percentuale (46%).

	METRICA	ADULTI [0,5343]	GIOVANI [0,4601]	TOTALE
Best Lift e Recall globale	Recall	0,7864	0,3077	-
	Accuracy	0,5711	0,5538	-
	Utenti Assegnati	1886	651	2537
	Utenti Non Assegnati	0	0	0
	Regole Generate	183	85	268
	Recall Globale = 0,5654 confidenza = 0,51		Lift = 1,0506 supporto = 35	

Tabella 6.19: *dataset Yahoo! - Risultati per la caratterizzazione degli utenti in base all'età, applicando il metodo basato sulle AR. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

Singular Value Decomposition

In questa modalità, a differenza di quanto avvenuto per i test relativi all'inferenza del sesso, il Lifestyle degli utenti è stato rappresentato utilizzando una matrice formata da un numero di colonne pari al numero delle classi (vedi Paragrafo 5.2.1).

Scegliendo dei valori di *k* e *t* pari rispettivamente a 6 e $-0{,}5$ si ottengono sia il miglior Lift, pari a $1{,}0212$, che la migliore Recall globale, il cui valore è pari a $0{,}5473$. La configurazione che consente di ottenere tali risultati è riportata nella Tabella 6.20.

	METRICA	ADULTI [0,5643]	GIOVANI [0,4601]	TOTALE
Best Lift e Recall globale	Recall	0,5940	0,4934	-
	Accuracy	0,5752	0,5128	-
	Utenti Assegnati	1264	1020	2284
	Utenti Non Assegnati	0	0	0
	Recall Globale = 0,5473 $k_{best} = 6$		Lift = 1,0212 $t_{best} = -0{,}5$	

Tabella 6.20: *dataset Yahoo! - Risultati per la caratterizzazione degli utenti in base all'età, applicando il metodo basato su SVD. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

In questa configurazione per tutti gli utenti viene proposta una predizione sulla classe di appartenenza, quindi i valori di Accuracy e Recall globali coin-

cidono. Sia per quanto riguarda la classe ADULTI che per la classe GIOVANI, i valori di Accuracy e Recall sono superiori alla percentuale di utenti che compongono la classe.

Support Vector Machines

I risultati dei test effettuati con la modalità *C*-classification sono riportati nella Tabella 6.21.

Matrice Utilizzata	Kernel	Scaling	C	γ	Lift	Recall globale	Accuracy globale	Classi assegnate
UGMnorm	Lineare	-	8192	-	1,0084	0,5744	0,5744	2 su 2
	Lineare	V [0, 1]	32	-	1,0008	0,5639	0,5639	2 su 2
	Lineare	A[0, 1]	512	-	1,0000	0,5897	0,5897	2 su 2
	Lineare	V[−1, 1]	32768	-	1,0047	0,5661	0,5661	2 su 2
	Lineare	A[−1, 1]	32768	-	1,0000	0,5867	0,5867	2 su 2
	Radiale	-	32768	8	1,0000	0,5625	0,5625	2 su 2
	Radiale	A[0,1]	8192	2	1,0337	0,5770	0,5770	2 su 2
	Radiale	A[−1,1]	8192	8	1,0282	0,5739	0,5739	2 su 2
UGMnoNorm	Lineare	A[0, 1]	32768	-	1,0084	0,5805	0,5805	2 su 2
	Lineare	A[−1, 1]	32768	-	1,0198	0,5871	0,5871	2 su 2
	Radiale	A[0,1]	32768	8	1,0074	0,5941	0,5941	2 su 2
	Radiale	A[−1,1]	32768	8	1,0230	0,5844	0,5844	2 su 2

Tabella 6.21: *Inferenza dell'età sul dataset Yahoo! attraverso l'utilizzo di una SVM in modalità C-classification. Le lettere 'V' e 'A' indicano rispettivamente l'utilizzo dello scaling solo durante la procedura di Cross Validation e il caso in cui esso viene applicato anche nella fase di apprendimento e test.*

Il miglior risultato in termini di Recall globale ($0{,}5941$) si ottiene facendo riferimento ad un kernel di tipo radiale ed applicando uno scaling totale nell'intervallo $(0, 1)$ alla matrice UGM non normalizzata, con C e γ che assumono rispettivamente i valori 32768 e 8. Il miglior risultato in termini di Lift ($1{,}0337$) si ottiene invece facendo riferimento ad un kernel di tipo radiale ed applicando uno scaling totale nell'intervallo $(0, 1)$, con C e γ pari rispettivamente a 8192 e 2; I valori relativi alle Accuracy e alle Recall delle singole classi per le due configurazioni appena descritte sono mostrate nella Tabella 6.22.
Basandosi sulla modalità *Nu*-classification si ottengono i risultati espressi nella Tabella 6.23.

Il miglior risultato in termini di Recall globale ($0{,}5739$) si ottiene facendo riferimento ad un kernel di tipo lineare, senza applicare lo scaling ed effettuando la normalizzazione della matrice UGM. La combinazione di parametri che con-

	METRICA	GIOVANI [0,4610]	ADULTI [0,5390]
Best Lift	Accuracy	0,5137	0,6093
	Recall	0,1994	0,8687
	Recall globale = 0,5770		Lift = 1,0337
	C = 8192 γ = 2 UGM normalizzata Kernel: RADIALE Scaling: [0, 1]		
Best Recall	Accuracy	0,5623	0,5796
	Recall	0,1929	0,8817
	Recall globale = 0,5941		Lift = 1,0074
	C = 32768 γ = 8 UGM non normalizzata Kernel: RADIALE Scaling: [0, 1]		

Tabella 6.22: *dataset Yahoo! - Risultati per la caratterizzazione degli utenti in base all'età, applicando il metodo basato sul SVM in modalità C-classification.*

Matrice Utilizzata	Kernel	Scaling	C	γ	ν	Lift	Recall globale	Accuracy globale	Classi assegnate
UGMnorm	Lineare	-	0,125	-	0,80	1,0077	0,5739	0,5739	2 su 2
	Lineare	A[0, 1]	0,125	-	0,60	0,7055	0,4160	0,4160	2 su 2
	Lineare	V[0, 1]	0,125	-	0,60	0,6981	0,4116	0,4116	2 su 2
	Lineare	A[−1, 1]	0,125	-	0,50	0,9681	0,5709	0,5709	2 su 2
	Lineare	V[−1, 1]	0,125	-	0,50	0,7782	0,4589	0,4589	2 su 2
	Radiale	-	0,125	0,0078	0,12	0,9565	0,5477	0,5477	2 su 2
	Radiale	A[0, 1]	0,125	2^{-11}	0,61	0,7003	0,4130	0,4130	2 su 2
	Radiale	A[−1, 1]	0,125	2^{-13}	0,61	0,6988	0,4121	0,4121	2 su 2
UGMnoNorm	Lineare	A[0, 1]	0,125	-	0,50	0,8613	0,5079	0,5079	2 su 2
	Lineare	A[−1, 1]	0,125	-	0,60	0,8353	0,4926	0,4926	2 su 2
	Radiale	A[0, 1]	0,125	8	0,59	0,7611	0,4488	0,4488	2 su 2
	Radiale	A[−1, 1]	0,125	8	0,74	0,9169	0,5407	0,5407	2 su 2

Tabella 6.23: *Inferenza dell'età sul dataset Yahoo! attraverso l'utilizzo di una SVM in modalità Nu-classification. Le lettere 'V' e 'A' indicano rispettivamente l'utilizzo dello scaling solo durante la procedura di Cross Validation e il caso in cui esso viene applicato anche nella fase di apprendimento e test.*

sente di ottenere tali risultati è mostrata nella Tabella 6.24. Tale configurazione è anche la migliore in termini di Lift per la categoria *Nu*-classification.

	METRICA	GIOVANI [0,4610]	ADULTI [0,5390]
Best Lift e Recall globale	Accuracy	0,5073	0,6025
	Recall	0,3537	0,7404
	Recall globale = 0,5739		Lift = 1,0077
	C = 0,125 ν = 0,80 UGM normalizzata Kernel: LINEARE Scaling: -		

Tabella 6.24: *dataset Yahoo! - Risultati per la caratterizzazione degli utenti in base all'età, applicando il metodo basato su SVM in modalità Nu-classification. Con questa configurazione di parametri si ottengono sia il miglior Lift che la migliore Recall globale.*

Test con minore numero di generi

Analogamente a quanto visto per il dataset MovieLens, i risultati dei test effettuati eliminando i rating associati ai generi di film ritenuti meno significativi dall'utente dalla matrice UGMnorm sono inferiori rispetto alle configurazioni che garantiscono le migliori performance descritte in questo capitolo. Utilizzando ad esempio la modalità *C*-classification, uno scaling compreso tra 0 e 1 ed un kernel di tipo radiale con un numero massimo di generi preferiti pari a 3, otteniamo una Recall globale uguale a $0{,}70954$ ed un Lift pari a $0{,}9878$, risultati quindi decisamente inferiori rispetto alla versione standard del test con gli stessi settaggi.

6.3 Confronto dei risultati ottenuti

In questo paragrafo si effettua un confronto tra i migliori risultati ottenuti per ciascuna tipologia di test, con l'obiettivo di individuare l'algoritmo di base che fornisca le migliori prestazioni per ciascuna tipologia in esame.

Nella prima sezione sono riportate le considerazioni effettuate analizzando i risultati relativi all'inferenza del sesso, facendo riferimento dapprima ai test eseguiti sul dataset MovieLens ed in seguito ai corrispettivi test effettuati sul

dataset Yahoo!; nel Paragrafo 6.3.2, invece, sono mostrati i migliori risultati ottenuti nei test sull'inferenza dell'età degli utenti, prima relativamente al dataset MovieLens, poi al dataset Yahoo!.

6.3.1 Inferenza del sesso

Per quanto riguarda i test sull'inferenza del sesso degli utenti del dataset MovieLens, i migliori risultati vengono ottenuti applicando la soluzione basata su SVM, che determina un Lift pari a $1{,}0821$ ed una Recall complessiva pari a $0{,}7925$, entrambe superiori ai corrispettivi valori ricavati applicando le metodologie che utilizzano AR ed SVD.

Anche dal punto di vista delle singole classi, SVM fornisce risultati migliori rispetto a quelli ottenuti applicando gli altri due algoritmi di classificazione; l'unica metrica per cui l'applicazione di SVM non garantisce il miglior risultato è la Recall della classe FEMMINE, per cui SVD consente di ottenere un valore pari a $0{,}4765$, superiore al valore $0{,}3897$ registrato per i test basati su SVM. La Tabella 6.27 mostra i migliori risultati ottenuti per ciascuna delle tre soluzioni implementate.

METRICA	AR	SVD	SVM
Lift	1,0440	1,0484	1,0821
Recall Globale	0,7488	0,7533	0,7925
Recall MASCHI	0,9341	0,8617	0,9397
Recall FEMMINE	0,2790	0,4765	0,3897
Accuracy Globale	0,7488	0,7533	0,7925
Accuracy MASCHI	0,7665	0,8078	0,8082
Accuracy FEMMINE	0,6265	0,5745	0,7026

Tabella 6.25: *confronto tra i migliori risultati in termini di Recall globale, ottenuti nei test di inferenza del sesso degli utenti appartenenti al dataset MovieLens applicando i tre algoritmi di base.*

I risultati per l'inferenza del sesso degli utenti del dataset Yahoo! sono globalmente inferiori rispetto a quelli relativi al dataset MovieLens riportati in Tabella 6.25. In questo caso i risultati ottenuti per le tre metodologie di classificazione sono simili, contrariamente a quanto avveniva relativamente a MovieLens, in cui la soluzione basata su SVM garantiva risultati migliori per tutte le metriche di valutazione. Anche per questa tipologia di test SVM

consente di ottenere i risultati migliori a livello globale, ma il divario con AR ed SVD è molto limitato.

Per quanto riguarda le Recall delle singole classi, le migliori performance per la classe MASCHI si ottengono applicando le Regole di Associazione, ottenendo un valore pari a $0{,}9516$, mentre il migliore risultato per la classe FEMMINE ($0{,}2697$) è ricavato dall'applicazione di SVD. Utilizzando SVM, invece, si ottengono risultati leggermente inferiori ai due migliori, ma la combinazione risulta essere la più equilibrata, essendo i due valori di Recall così ottenuti pari a $0{,}9330$ e $0{,}2610$.

METRICA	AR	SVD	SVM
Lift	1,0271	1,0201	1,0387
Recall Globale	0,7328	0,7314	0,7393
Recall MASCHI	0,9516	0,9136	0,9330
Recall FEMMINE	0,1880	0,2697	0,2610
Accuracy Globale	0,7328	0,7314	0,7393
Accuracy MASCHI	0,7448	0,7612	0,7571
Accuracy FEMMINE	0,6093	0,5521	0,6121

Tabella 6.26: *confronto tra i migliori risultati in termini di Recall globale, ottenuti nei test di inferenza del sesso degli utenti appartenenti al dataset Yahoo! applicando i tre algoritmi di base.*

6.3.2 Inferenza dell'età

Per quanto riguarda il dataset MovieLens l'applicazione di SVM consente di ricavare i migliori risultati per tutte le metriche di valutazione, ad eccezione della Recall associata alla classe ADULTI. In particolare dall'applicazione di tale metodologia si ottengono valori di Lift e Recall complessiva pari rispettivamente a $1{,}1305$ e $0{,}6071$. Inoltre quello basato su SVM è l'unico metodo che consente di assegnare degli utenti alla classe GIOVANISSIMI, anche se il valore di Recall ottenuto è pari a $0{,}0250$; il valore di Accuracy corripondente è pari ad uno, segno che l'algoritmo ha correttamente classificato tutti gli utenti individuati per tale classe.

Per il dataset Yahoo! la migliore configurazione si ottiene suddividendo gli utenti in due classi di età: GIOVANI ed ADULTI; non è quindi possibile effettuare un confronto con il dataset MovieLens, visto che i migliori risultati

METRICA	AR	SVD	SVM
Lift	1,0710	1,0401	1,1305
Recall Globale	0,5672	0,5419	0,6071
Recall GIOVANISSIMI	–	–	0,0250
Recall GIOVANI	0,8371	0,6896	0,8468
Recall ADULTI	0,3342	0,5182	0,4164
Recall ANZIANI	–	–	–
Accuracy Globale	0,5672	0,5419	0,6071
Accuracy GIOVANISSIMI	–	–	1,0000
Accuracy GIOVANI	0,5873	0,6123	0,6257
Accuracy ADULTI	0,5054	0,4439	0,5558
Accuracy ANZIANI	–	–	–

Tabella 6.27: *confronto tra i migliori risultati in termini di Recall globale, ottenuti nei test di inferenza dell'età degli utenti appartenenti al dataset MovieLens applicando i tre algoritmi di base.*

per tale base di dati si ricavano suddividendo gli utenti in quattro classi. Il miglior Lift si registra applicando il metodo delle Regole di Associazione ed è pari a $1{,}0506$, mentre la migliore Recall è pari a $0{,}5941$ e si ottiene con il metodo basato su SVM. Come già osservato nel caso dei test sulla caratterizzazione del sesso degli utenti, nel dataset Yahoo! la superiorità di SVM è meno marcata rispetto a quanto verificato per MovieLens. Analizzando i dati relativi alle Recall delle due classi di età, tuttavia, si può notare come i valori riscontrati per SVM siano nettamente più equilibrati rispetto alle altre due metodologie, essendo pari $0{,}5623$ a per la classe GIOVANI e a $0{,}5796$ per la classe ADULTI.

In conclusione è possibile affermare che la soluzione basata su SVM si mostra senza dubbio la più performante tra le tre implementate per quanto riguarda il dataset MovieLens, consentendo di ottenere i migliori risultati sia per le metriche globali, che relativamente alle singole classi.

I risultati ottenuti per il dataset Yahoo!, invece, si dimostrano essere più simili a livello globale, anche se l'applicazione di SVM fornisce valori più equilibrati tra le Recall delle varie classi, e risulta quindi essere preferibile rispetto all'utilizzo di AR ed SVD.

METRICA	AR	SVD	SVM
Lift	1,0506	1,0212	1,0074
Recall Globale	0,5654	0,5473	0,5941
Recall GIOVANI	0,3077	0,4934	0,5623
Recall ADULTI	0,7864	0,5940	0,5796
Accuracy Globale	0,5654	0,5473	0,5941
Accuracy GIOVANI	0,5538	0,5128	0,1921
Accuracy ADULTI	0,5711	0,5752	0,8817

Tabella 6.28: *confronto tra i migliori risultati in termini di Recall globale, ottenuti nei test di inferenza dell'età degli utenti appartenenti al dataset Yahoo! applicando i tre algoritmi di base.*

Capitolo 7

Combinazione di classificatori

Indice

In questo capitolo vengono presentate le soluzioni implementate per la combinazione delle predizioni fornite dagli algoritmi di base relativamente al Lifestyle degli utenti. Dopo aver descritto nel Paragrafo 7.1 gli schemi implementativi seguiti, vengono mostrati i risultati per entrambe le soluzioni implementate (Paragrafo 7.2); il confronto con i risultati ottenuti applicando separatamente i singoli algoritmi è invece presentato nel Paragrafo 7.3.

7.1 Soluzioni combinate per l'inferenza del Lifestyle

Dall'analisi dei risultati ottenuti applicando separatamente le tre metodologie operative descritte nel Capitolo 5 è emerso come, nonostante i risultati complessivi possano essere ritenuti soddisfacenti, tutti i classificatori da noi

considerati tendano ad assegnare gli utenti alle classi più popolate, a scapito delle classi con minor numero di utenti nell'insieme di apprendimento.
La seconda fase del nostro lavoro è stata dunque svolta in modo da realizzare i seguenti obiettivi:

1. ottenere un miglioramento delle performance complessive della fase di classificazione degli utenti;
2. costruire un modello in grado di assegnare gli utenti anche alle classi di dimensioni minori.

Come proposto da Granara in [19], una possibile soluzione per tentare di eliminare il problema della mancata assegnazione degli utenti alle classi di dimensioni minori consiste nella costruzione di un modello a partire da un insieme di apprendimento bilanciato, in cui tutte le classi siano costituite dallo stesso numero di utenti.

Per costruire un insieme bilanciato è stato necessario in primo luogo individuare la classe formata dal minor numero di utenti. Successivamente, per ognuna delle altre classi previste, è stato estratto dall'insieme di apprendimento un numero di utenti pari a quello della classe di dimensione minore; tale selezione è stata effettuata in modo casuale, per evitare di utilizzare dati relativi allo stesso periodo temporale.

In seguito, a partire dall'insieme di apprendimento bilanciato così ricavato, si è provveduto alla costruzione del modello per ciascuno dei tre metodi di base, secondo le modalità descritte in precedenza nel Capitolo 5. I modelli ricavati sono stati utilizzati per la classificazione degli utenti appartenenti all'insieme di test, costruito selezionando casualmente il 30% degli utenti del dataset, sul quale non sono quindi state compiute operazioni di bilanciamento. La Tabella 7.1 mostra il confronto tra i migliori risultati ottenuti applicando SVM in versione tradizionale e bilanciata, per effettuare test sull'inferenza dell'età degli utenti del dataset MovieLens.

Dall'analisi dei risultati ricavati applicando il bilanciamento dei dataset è possibile notare come soltanto uno degli obiettivi che ci siamo posti sia stato effettivamente raggiunto, dato che i miglioramenti ottenuti dal punto di vista dell'assegnazione degli utenti alle classi minori coincidono con una diminuzione dei risultati complessivi in termini di Recall e Lift.

Dato che il ricorso al bilanciamento dell'insieme di apprendimento non ha fornito i risultati sperati, abbiamo scelto di implementare alcune tra le tecniche

METRICA	SVM TRADIZIONALE	SVM BILANCIATO
Recall Globale	0,6071	0,3642
Recall GIOVANISSIMI	0,0025	0,3226
Recall GIOVANI	0,8469	0,2611
Recall ADULTI	0,4164	0,5208
Recall ANZIANI	–	0,3486
Accuracy Globale	0,6071	0,3642
Accuracy GIOVANISSIMI	1,000	0,0678
Accuracy GIOVANI	0,6257	0,6919
Accuracy ADULTI	0,5558	0,4066
Accuracy ANZIANI	–	0,1324
Lift	1,1305	0,5200

Tabella 7.1: *Dataset MovieLens: il confronto tra i risultati ottenuti dall'applicazione di SVM in versione tradizionale e bilanciata per l'inferenza dell'età degli utenti.*

basate sulla combinazione delle soluzioni di base descritte in precedenza nel Paragrafo 2.4, in modo da costruire un classificatore finale capace di centrare entrambi gli obiettivi prefissati.
Sono state implementate due soluzioni, entrambe basate sull'utilizzo di due tecniche combinate tra di loro:

1. il *Grading* e il *Weighted Voting*;
2. l'*Arbitro* e il *Weighted Voting*.

Per entrambe le soluzioni individuate sono state utilizzate le informazioni memorizzate nel dataset MovieLens, che consente di ottenere i migliori risultati nell'applicazione dei singoli classificatori di base.

7.1.1 Scelta dei classificatori di base

Il primo passo da effettuare nello sviluppo di questo tipo di soluzioni è la scelta di quali classificatori di base utilizzare nella fase di stima iniziale degli item. Nel nostro caso sono stati individuati quattro classificatori di base: il primo costruito basandosi sulle Regole di Associazione, il secondo implementato applicando la Singular Value Decomposition e gli ultimi due ricavati utilizzando la metodologia delle Support Vector Machines, scegliendo

le due configurazioni di parametri che hanno consentito di ottenere i migliori risultati nei test effettuati in precedenza.

Per ciascun algoritmo sono state anche elaborate le corrispondenti versioni bilanciate, in modo da ottenere un numero maggiore di classificatori di base e rendere l'analisi più completa, oltre a favorire l'assegnazione degli utenti alle classi meno popolate, sfruttando le peculiarità della tecnica del bilanciamento del dataset.

7.1.2 Grading e Weighted Voting

La prima metodologia sperimentata prevede che gli utenti dell'insieme di test vengano inizialmente assegnati ad una classe utilizzando separatamente i classificatori di base prescelti; in un secondo momento ciascuna delle stime fornite viene valutata attraverso un grader, in modo da eliminare quelle considerate non attendibili. L'utente è infine assegnato alla classe che ottiene il maggior numero di voti tra le stime dei classificatori di base valutate positivamente dal corrispettivo grader.
Presentiamo ora le fasi seguite per realizzare lo schema implementativo appena descritto:

Individuazione dei parametri per i classificatori di base

L'obiettivo della prima fase operativa è stato quello di individuare i migliori parametri per ciascun classificatore di base, da applicare in seguito al modello finale per la classificazione degli utenti. Dopo avere eseguito sul dataset tutte le operazioni preliminari già descritte per i singoli classificatori di base si è provveduto alla suddivisione degli utenti nei consueti insiemi di apprendimento, utilizzato per ricavare i modelli ed i migliori parametri richiesti dai vari metodi, e di test, su cui effettuare la classificazione finale degli utenti e calcolare le metriche di valutazione per le stime fornite. L'individuazione dei migliori parametri è stata effettuata applicando all'insieme di apprendimento la tecnica *10-folds cross validation*, già descritta in precedenza.

Individuazione dei parametri per il grader

L'insieme di apprendimento utilizzato nella fase precedente è stato suddiviso in due sottoinsiemi: il primo utilizzato per ricavare un modello per

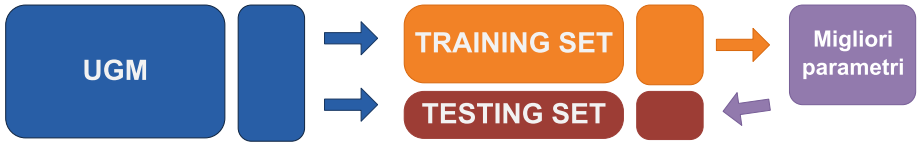


Figura 7.1: *individuazione dei migliori parametri per i classificatori di base, a partire dall'insieme di apprendimento.*

ciascun classificatore di base; il secondo, che è stato invece utilizzato per la costruzione di un insieme di apprendimento/validazione per il grader.

Gli utenti appartenenti al secondo sottoinsieme sono stati in un primo momento assegnati ad una classe utilizzando il modello appena ricavato a partire dall'analisi del primo sottoinsieme, al quale sono stati applicati i parametri migliori per la classificazione, individuati nella prima fase. Successivamente le stime ricavate dai classificatori di base sono state confrontate con le reali classi di appartenenza degli utenti, in modo da costruire le etichette da fornire al grader; tali etichette contengono il valore 1 nel caso in cui il classificatore abbia predetto correttamente la classe di appartenenza dell'utente mentre, in caso di errore, all'etichetta è stato assegnato un valore pari a -1. Le etichette contenenti la valutazione della classificazione sono quindi state concatenate ai feedback espressi dagli utenti, ottenendo l'insieme di apprendimento/validazione per il grader.

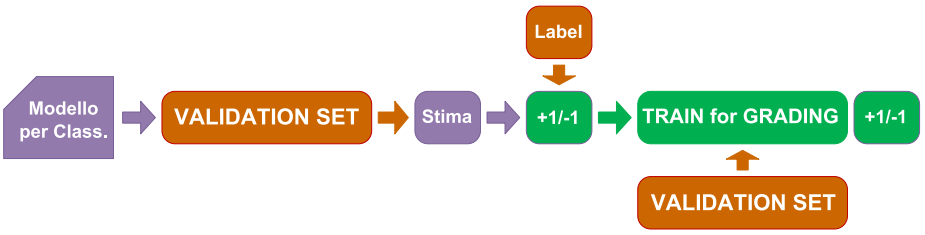


Figura 7.2: *il processo per la creazione dell'insieme di apprendimento del grader. Le stime delle classi di appartenenza degli utenti sono confrontate con i valori reali per creare le etichette da concatenare ai feedback relativi ai generi preferiti.*

Tale insieme è stato ulteriormente suddiviso in due partizioni, rispettivamente utilizzate per la costruzione di un modello per il grader e per l'individuazione dei corrispettivi migliori parametri. In questo caso si è ricorsi ad una

tecnica di cross-validation, senza operare una divisione in fold, per evitare di utilizzare insiemi di dimensione poco significativa rispetto al totale degli utenti a disposizione.

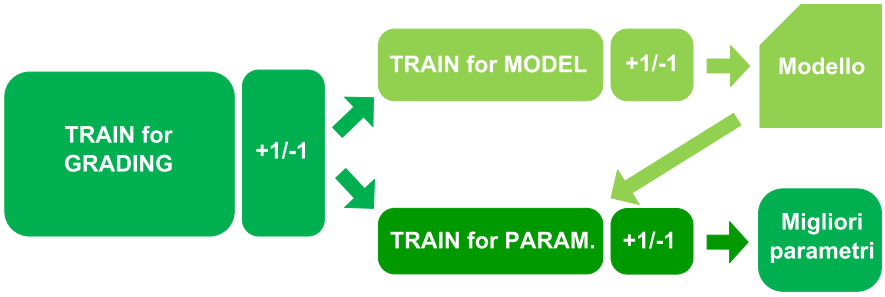


Figura 7.3: *l'insieme di apprendimento per il grader viene suddiviso in due sottoinsiemi, dedicati rispettivamente alla costruzione del modello e all'individuazione dei migliori parametri.*

Costruzione dei modelli per i classificatori di base e per il grader

In questa fase sono stati costruiti i modelli da utilizzare per la classificazione degli utenti dell'insieme di test e per la valutazione delle stime fornite dai classificatori di base da parte dei grader. Gli utenti dell'insieme di apprendimento iniziale sono stati nuovamente suddivisi in due sottoinsiemi: il primo è stato utilizzato per ricavare il modello finale per la classificazione degli utenti in base all'età o al sesso, mentre il secondo è stato utilizzato per la costruzione del modello per il grader, eseguendo le stesse operazioni svolte in precedenza.

Gli utenti sono stati stimati utilizzando il modello per la classificazione ed i corrispettivi migliori parametri; tali stime sono state confrontate con la classe reale di appartenenza ed infine sono state costruite le etichette contenenti il risultato della valutazione. Dalla concatenazione delle etichette con i vettori contenenti le preferenze degli utenti è stato creato l'insieme di apprendimento/valutazione finale per il grader, dal quale è stato ricavato il modello da applicare agli utenti dell'insieme di test.

Applicazione dei modelli e assegnazione della classe

I due modelli così costruiti ed i corrispettivi migliori parametri sono stati quindi utilizzati separatamente per predire la classe di appartenenza di ciascuno degli utenti di test e per valutare l'attendibilità di tale predizione. La classificazione finale degli utenti è stata realizzata utilizzando la tecnica del Weighted Voting, assegnando un peso solo alle predizioni dei classificatori di base considerate attendibili. A seconda del tipo di classificazione da effettuare, in base il sesso o l'età, sono stati assegnati pesi differenti a ciascun classificatore di base.

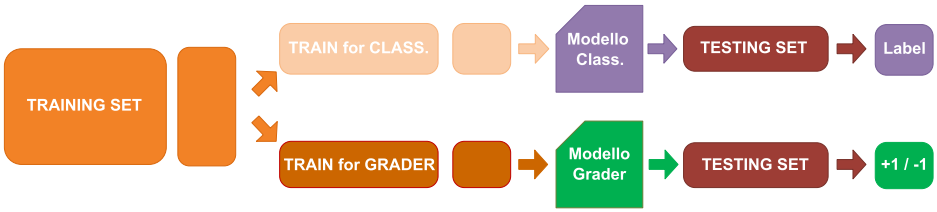


Figura 7.4: *per gli utenti dell'insieme di test si applicano i due modelli ricavati in precedenza, ottenendo una stima della classe di appartenenza e una valutazione sull'attendibilità della previsione effettuata.*

7.1.3 Arbitro e Weighted Voting

Lo schema implementativo per questa soluzione è molto simile a quello previsto per la metodologia che fa uso di Grading e Weighted Voting; la principale differenza consiste nel fatto che, mentre nel caso precedente per tutti gli utenti venivano creati due modelli, utilizzati rispettivamente per la stima della classe di appartenenza e per la valutazione di tale stima da parte del grader, in questa metodologia vengono sottoposti all'arbitro solo gli utenti per i quali si verificano discordanze nelle previsioni della classe di appartenenza fornite dai vari classificatori di base.
La Selection Rule utilizzata è quindi la seguente:

l'insieme di apprendimento dell'arbitro è formato dai record associati agli utenti per i quali si verificano contrasti nelle stime prodotte dai classificatori di base.

Se tutti i classificatori di base sono concordi nell'assegnare un utente ad una determinata classe non è necessario l'utilizzo dell'arbitro, in quanto la stima fornita dai classificatori di base viene considerata sufficientemente affidabile.

Un'altra differenza fondamentale risiede nel numero di classificatori di base utilizzati che, in questo caso, è pari a quattro. Tale ruolo è stato infatti assegnato ai metodi basati su Regole di Associazione e SVD, in versione tradizionale e bilanciata, mentre la funzione di arbitro viene svolta da SVM, che risulta essere l'algoritmo più accurato tra i tre analizzati.

Individuazione dei parametri e dei modelli per i classificatori di base e per l'arbitro

Il dataset iniziale è stato suddiviso nelle stesse partizioni create per l'implementazione della prima soluzione; anche le operazioni eseguite per l'individuazione dei parametri e la costruzione del modello per i classificatori di base sono state le medesime descritte in precedenza.

Per quanto riguarda l'arbitro, invece, è stata applicata la Selection Rule sulle relative partizioni del dataset, in modo da costruire gli insiemi di apprendimento/validazione da cui ricavare i parametri migliori e il modello a cui sottoporre gli utenti dell'insieme di test.

Applicazione dei modelli e assegnazione della classe

Nell'ultima fase sono stati applicati i modelli per la classificazione agli utenti appartenenti all'insieme di test, in modo da fornire per ciascuno un numero di stime pari al numero dei classificatori di base. Gli utenti per cui si è verificata l'unanimità nelle stime di base sono stati assegnati alla classe indicata, mentre gli altri sono stati ulteriormente classificati, utilizzando il modello ed i parametri migliori ricavati per l'arbitro nelle fasi precedenti. La stima così ricavata è stata combinata con quelle fornite in precedenza dai classificatori di base mediante la metodologia del Weighted Voting, in modo da ricavare la classe a cui assegnare gli utenti per cui si erano registrati conflitti nella prima fase di classificazione.

7.2 Risultati

In questo paragrafo sono mostrati i risultati dell'applicazione delle due metodologie di combinazione di classificatori implementate, la prima ottenuta ricorrendo alle tecniche di Grading e Weighted Voting, la seconda che prevede l'utilizzo delle tecniche di Arbitro e Weighted Voting.

7.2.1 Grading e Weighted Voting

Come illustrato nel Paragrafo 7.1.2, tale soluzione è suddivisa in due fasi: nella prima si producono per ciascun utente le stime dei classificatori di base e dei relativi grader, nella seconda si assegnano gli utenti alle classi utilizzando la tecnica del Voting. Mentre la prima fase è comune per i test basati sul sesso e sull'età, la seconda è stata implementata in due modi differenti a seconda del tipo di test effettuato.

Risultati relativi all'inferenza del sesso

Nei test effettuati per l'inferenza del sesso degli utenti, la votazione è stata implementata calcolando un punteggio per ciascun utente dell'insieme di test secondo le seguenti regole:

- se l'algoritmo assegna l'utente alla classe MASCHIO e il grader valuta positivamente la previsione effettuata, il punteggio associato all'utente in esame viene incrementato di 1;
- se l'algoritmo assegna l'utente alla classe FEMMINA e il grader valuta positivamente la previsione effettuata, il punteggio associato all'utente in esame viene decrementato di 1;
- se il grader valuta in modo negativo la previsione, indipendentemente dalla classe predetta, il punteggio associato all'utente in esame non viene modificato;
- se l'algoritmo di base non assegna l'utene ad alcuna classe, il punteggio associato all'utente in esame non viene modificato.

Un punteggio totale positivo, di conseguenza, indica che la maggioranza dei metodi di base per i quali il corrispondente grader ha valutato correttamente la previsione ha assegnato l'utente alla classe MASCHIO; l'ottenimento

di un punteggio negativo, invece, indica che la maggioranza dei metodi ha assegnato l'utente alla classe FEMMINA.

A partire dai risultati ottenuti utilizzando la configurazione di base sono in seguito stati effettuati dei test modificando il valore di soglia utilizzato come discriminante nell'assegnazione degli utenti alle due classi. L'utilizzo di un valore di soglia inferiore allo zero indica che la maggioranza richiesta per l'assegnazione di un utente alla classe FEMMINE è superiore rispetto a quella necessaria nel caso base, con soglia uguale a zero. L'utilizzo di valori di soglia maggiori di zero, invece, indica l'uso di criteri più restrittivi per l'assegnazione degli utenti alla classe MASCHI.

Successivamente sono stati eseguiti test con l'obiettivo di verificare l'utilità dell'utilizzo della tecnica del Grading: sono state sperimentate varie possibilità, nelle quali tale tecnica è stata applicata soltanto ad alcuni sottoinsiemi dei classificatori di base. L'unica configurazione che ha consentito di ottenere risultati significativi è quella in cui non si applica il Grading ai classificatori di base che utilizzano SVM.

Per completare l'analisi si è quindi provveduto ad effettuare una serie di test nei quali si sono modificati di volta in volta i pesi associati ai singoli classificatori di base, con l'obiettivo di individuare una configurazione che consentisse di ottenere le migliori performance.

La Tabella 7.2 mostra la configurazione per cui si ottengono i migliori risultati in termini assoluti, ottenuta senza applicare la tecnica del Grading ai classificatori che utilizzano SVD. In questo caso sono stati raddoppiati i pesi associati alla classe MASCHI per quanto riguarda i metodi tradizionali, e FEMMINE per quanto riguarda i classificatori bilanciati. La soglia applicata è uguale a zero.

I risultati più equilibrati, invece, si ottengono applicando un grader a tutti i classificatori di base (Tabella 7.3). In questo caso sono stati raddoppiati i pesi associati alla classi FEMMINE per quanto riguarda i metodi tradizionali, e MASCHI per quanto riguarda i classificatori bilanciati; la soglia applicata è pari ad uno.

Risultati relativi all'inferenza dell'età

In questa tipologia di test sono stati calcolati per ciascun utente quattro punteggi, ognuno corrispondente ad una delle classi previste, assegnando

METRICA	Soglia = 0
Recall Globale	0,7638
Recall MASCHI	0,9284
Recall FEMMINE	0,3624
Accuracy Globale	0,7638
Accuracy MASCHI	0,7803
Accuracy FEMMINE	0,6749
Lift	1,0770

Tabella 7.2: *i migliori risultati ottenuti applicando le tecniche di Grading e Weighted Voting per l'inferenza del sesso degli utenti del dataset MovieLens. La configurazione prevede di applicare un grader a tutti i classificatori di base ed assegnare ad ognuno di essi un peso unitario.*

METRICA	Soglia = 1
Recall Globale	0,7434
Recall MASCHI	0,8234
Recall FEMMINE	0,5484
Accuracy Globale	0,7434
Accuracy MASCHI	0,8164
Accuracy FEMMINE	0,5601
Lift	1,0482

Tabella 7.3: *i risultati più equilibrati ottenuti applicando le tecniche di Grading e Weighted Voting per l'inferenza del sesso degli utenti del dataset MovieLens, ottenuti applicando un grader a tutti i classificatori di base e raddoppiando il peso associato alla classe FEMMINE per i metodi tradizionali e MASCHI per i metodi bilanciati.*

METRICA	VALORE
Recall Globale	0,5982
Recall GIOVANISSIMI	0,0448
Recall GIOVANI	0,8784
Recall ADULTI	0,3338
Recall ANZIANI	0,0182
Accuracy Globale	0,5982
Accuracy GIOVANISSIMI	0,3000
Accuracy GIOVANI	0,6099
Accuracy ADULTI	0,5871
Accuracy ANZIANI	0,1053
Lift	1,1073

Tabella 7.4: *i migliori risultati ottenuti applicando le tecniche di Grading e Weighted Voting per l'inferenza del sesso degli utenti del dataset MovieLens, ottenuti applicando un grader a tutti i classificatori di base e raddoppiando il peso associato alle classi GIOVANI ed ADULTI per i metodi tradizionali e GIOVANISSIMI ed ANZIANI per i metodi bilanciati*

l'utente alla classe per la quale è stato ottenuto il punteggio maggiore. Rispetto ai test basati sull'inferenza del sesso, quindi, non sono previsti valori di soglia per l'assegnazione alle classi; anche per questa tipologia di test sono state sperimentate configurazioni di pesi differenti

La Tabella 7.4 mostra la configurazione che consente di ottenere il miglior Lift, pari a 1,10730. In questa configurazione si conferisce un peso doppio ai voti relativi alle classi GIOVANI ed ADULTI per quanto riguarda i metodi non bilanciati; relativamente ai metodi bilanciati, invece, si è scelto di privilegiare i voti assegnati alle classi estreme, in modo da ottenere una distribuzione equilibrata, ma che non andasse troppo ad impattare sulle performance relative alle classi più popolate

La Tabella 7.5 mostra i risultati ottenuti mantenendo immutati i pesi associati alla classe GIOVANI per tutti i classificatori e raddoppiando i pesi associati alle altre classi. Questa combinazione è stata sperimentata per limitare il peso dei voti alla classe GIOVANI; essendo tale classe la più popolata si registra una diminuzione dei valori complessivi, al quale corrisponde però l'ottenimento di buone performnce per le altre classi previste.

METRICA	VALORI
Recall Globale	0,4752
Recall GIOVANISSIMI	0,3731
Recall GIOVANI	0,5240
Recall ADULTI	0,4253
Recall ANZIANI	0,4000
Accuracy Globale	0,4752
Accuracy GIOVANISSIMI	0,1101
Accuracy GIOVANI	0,6895
Accuracy ADULTI	0,4745
Accuracy ANZIANI	0,1739
Lift	0,8795

Tabella 7.5: *i risultati più equilibrati ottenuti applicando le tecniche di Grading e Weighted Voting per l'inferenza dell'età degli utenti del dataset MovieLens, ottenuti applicando un grader a tutti i classificatori di base e raddoppiando il peso associato a tutte le classi ad eccezione della classe GIOVANI.*

7.2.2 Arbitro e Weighted Voting

Anche questa metodologia prevede la realizzazione di due fasi implementative: inizialmente si effettua la stima del Lifestyle degli utenti applicando i classificatori di base; in seguito, per tutti gli utenti per i quali non si è raggiunta l'unanimità delle predizioni fornite nella prima fase, si costruisce un ulteriore classificatore, in grado di risolvere i conflitti registrati in precedenza. In questo paragrafo vengono mostrati in primo luogo i risultati dei test per l'inferenza del sesso, successivamente quelli dell'età, entrambi eseguiti sul dataset MovieLens.

Risultati relativi all'inferenza del sesso

In questa tipologia si utilizzano quattro classificatori di base: due basati sulle Regole di Associazione e due su SVD, sia in versione bilanciata che non.

Per quanto riguarda la fase di Voting si calcola un punteggio per ciascun utente, che viene determinato secondo le seguenti regole:

1. se l'arbitro assegna l'utente alla classe MASCHI, il punteggio viene incrementato di due;

2. se l'arbitro assegna l'utente alla classe FEMMINE, il punteggio viene decrementato di due;
3. se un classificatore di base assegna l'utente alla classe MASCHI, il punteggio viene incrementato di uno;
4. se un classificatore di base assegna l'utente alla classe FEMMINE, il punteggio viene decrementato di uno;
5. se un classificatore di base non assegna l'utente ad alcuna classe, il punteggio non viene modificato.

A differenza della soluzione basata sul Grading sono state utilizzate due soglie per l'assegnazione degli utenti alle classi: se il punteggio ottenuto da un utente è risultato inferiore alla prima soglia l'utente è stato assegnato alla classe FEMMINE mentre gli utenti con punteggi maggiori rispetto alla seconda soglia sono stati assegnati alla classe FEMMINE. Agli utenti il cui punteggio è compreso tra le due soglie è stata assegnata la classe predetta dall'arbitro, che si occupa quindi di giudicare le situazioni di maggiore incertezza.

Pur avendo sperimentato un numero significativo di combinazioni di pesi assegnate ai vari classificatori, le migliori performance sia in termini assoluti che in termini di equilibrio tra le varie classi si ottengono per quella di base, in cui tutti i classificatori hanno peso unitario; la Tabella 7.6 mostra i risultati ottenuti per tale configurazione.

Stabilendo un intervallo di incertezza maggiore, cioè dando maggiore importanza all'arbitro, si ottiene una diminuzione dei valori associati alle metriche globali, ma si può notare un miglioramento della Recall della classe FEMMINE.

Risultati relativi all'inferenza dell'età

Per i test relativi all'inferenza dell'età si utilizzano solamente due classificatori nella prima fase: uno basato su SVD in versione bilanciata, l'altro sulle Regole di Associazione, anch'esso in versione bilanciata. Tale scelta è stata motivata dal fatto che i classificatori non bilanciati non assegnano utenti alle classi meno popolate; di conseguenza il loro utilizzo nella fase iniziale favorirebbe l'assegnazione degli utenti alle classi più popolate.

Nella fase di Voting si è seguito il medesimo schema previsto per la soluzione basata sul Grading calcolando quattro punteggi per ciascun utente, ed

METRICA	SogliaM =1 SogliaF =-1	SogliaM = 3 SogliaF = -3
Recall Globale	0,7483	0,7185
Recall MASCHI	0,8920	0,8210
Recall FEMMINE	0,3586	0,4451
Accuracy Globale	0,7483	0,7185
Accuracy MASCHI	0,7905	0,7993
Accuracy FEMMINE	0,5503	0,4759
Lift	1,0242	0,9777

Tabella 7.6: *risultati relativi all'inferenza del sesso degli utenti del dataset MovieLens applicando le tecniche di Arbitro e Weighted Voting. La prima colonna mostra i risultati migliori in termini di Lift, ottenuti assegnando peso unitario e ricorrendo a soglie pari ad 1 e -1. La seconda colonna mostra i risultati più equilibrati tra quelli ottenuti, in cui le soglie utilizzate sono 3 e -3, mentre la distribuzione dei pesi rimane immutata.*

assegnandolo alla classe per la quale è stato ottenuto il punteggio più alto; in questa fase sono stati utilizzati tutti i classificatori, sia in versione bilanciata che non bilanciata. Anche in questo caso la migliore configurazione in termini assoluti si ottiene applicando un peso unitario a tutti i classificatori; i risultati per tale configurazione sono mostrati nella Tabella 7.7. Assegnando un peso doppio ai voti associati ai classificatori non bilanciati,invece, si ottiene la configurazione mostrata in Tabella 7.8 che risulta essere quella maggiormente bilanciata per questa tipologia di test.

METRICA	VALORE
Recall Globale	0,4393
Recall GIOVANISSIMI	0,1765
Recall GIOVANI	0,5768
Recall ADULTI	0,2866
Recall ANZIANI	0,3273
Accuracy Globale	0,4393
Accuracy GIOVANISSIMI	0,0916
Accuracy GIOVANI	0,6057
Accuracy ADULTI	0,4444
Accuracy ANZIANI	0,1088
Lift	0,8257

Tabella 7.7: *i migliori risultati relativi all'inferenza dell'età degli utenti del dataset MovieLens applicando le tecniche di Arbitro e Weighted Voting, ottenuti assegnando peso unitario a tutti i classificatori di base ed un peso doppio all'Arbitro.*

METRICA	VALORE
Recall Globale	0,4426
Recall GIOVANISSIMI	0,1765
Recall GIOVANI	0,5654
Recall ADULTI	0,3119
Recall ANZIANI	0,3273
Accuracy Globale	0,4426
Accuracy GIOVANISSIMI	0,0916
Accuracy GIOVANI	0,6165
Accuracy ADULTI	0,4485
Accuracy ANZIANI	0,1088
Lift	0,8320

Tabella 7.8: *i risultati più equilibrati relativi all'inferenza dell'età degli utenti del dataset MovieLens applicando le tecniche di Arbitro e Weighted Voting, ottenuti raddoppiando il peso associato ai classificatori bilanciati ed all'Arbitro.*

7.3 Confronto tra i risultati

In questo paragrafo si effettua un confronto tra i risultati ricavati applicando i singoli algoritmi di classificazione separatamente e combinandoli attraverso le due metodologie descritte in precedenza nel Paragrafo 7.1. Per quanto riguarda i classificatori di base, si propongono i risultati ottenuti con SVM, che si dimostrano i migliori. Relativamente alle due tipologie di caratterizzazione effettuate (in base al sesso e all'età) sono mostrate le configurazioni che assicurano le performance più elevate e quelle che forniscono i risultati più equilibrati tra le classi.

Inferenza del sesso degli utenti

La Tabella 7.9 mostra il confronto tra i migliori risultati ottenuti relativamente all'inferenza del sesso degli utenti. Si può osservare come, tra le due metodologie di classificazione proposte, le migliori performance siano ottenute applicando quella basata su Grading e Voting che, tuttavia, fornisce risultati complessivi inferiori rispetto all'utilizzo del solo SVM, in termini di Recall e Lift. I valori associati a quest'ultima metrica sono superiori ad uno per entrambe le soluzioni proposte .

Per quanto riguarda l'Accuracy, invece, entrambe le metodologie forniscono risultati superiori a SVM, soprattutto per quanto riguarda la classe FEMMINE, ad indicare una minore percentuale di errori per la classificazione di tali utenti. Questa osservazione è sicuramente positiva, dato che uno degli obiettivi della combinazione è di migliorare i risultati relativi alle classi con minore numero di utenti.

Nella Tabella 7.10, invece, sono mostrate le configurazioni che consentono di ottenere i risultati maggiormente equilibrati tra le varie classi; rispetto alle configurazioni contenute nella Tabella 7.9, è possibile notare una diminuzione dei Lift e delle Recall globali. Anche in questo caso la tecnica del Grading risulta migliore in confronto a quella dell'Arbitro. Rispetto alla configurazione maggiormente bilanciata ottenuta applicando SVM, le due metodologie di combinazione forniscono risultati meno equilibrati, privilegiando la classe MASCHI.

METRICA	Algoritmi di base (SVM)	Grading e Voting	Arbitro e Voting
Recall Globale	0,7925	0,7638	0,7483
Recall MASCHI	0,9397	0,9284	0,8920
Recall FEMMINE	0,3897	0,3624	0,3586
Accuracy Globale	0,7026	0,7638	0,7483
Accuracy MASCHI	0,8082	0,7803	0,7905
Accuracy FEMMINE	0,3897	0,6749	0,5503
Lift	1,0821	1,0770	1,0242

Tabella 7.9: *confronto tra i migliori risultati ottenuti nell'inferenza del sesso degli utenti, utilizzando i classificatori di base separatamente e combinandoli tra loro mediante le due soluzioni proposte.*

METRICA	Algoritmi di base (SVM)	Grading e Voting	Arbitro e Voting
Recall Globale	0,7412	0,7434	0,7185
Recall MASCHI	0,7477	0,8234	0,8210
Recall FEMMINE	0,7244	0,5484	0,4406
Accuracy Globale	0,7412	0,7434	0,7185
Accuracy MASCHI	0,8744	0,8164	0,7993
Accuracy FEMMINE	0,5280	0,5601	0,4759
Lift	1,0299	1,0482	0,9777

Tabella 7.10: *confronto tra i risultati più equilibrati ottenuti nell'inferenza del sesso degli utenti, utilizzando i classificatori di base separatamente e combinandoli tra loro mediante le due soluzioni proposte.*

METRICA	Algoritmi di base (SVM)	Grading e Voting	Arbitro e Voting
Recall Globale	0,6071	0,5982	0,4426
Recall Giovanissimi	0,0250	0,0448	0,1765
Recall Giovani	0,8460	0,8784	0,5654
Recall Adulti	0,4164	0,3338	0,3119
Recall Anziani	–	0,0182	0,3273
Accuracy Globale	0,6071	0,5982	0,4426
Accuracy Giovanissimi	1,0000	0,3000	0,0916
Accuracy Giovani	0,6257	0,6099	0,6165
Accuracy Adulti	0,5558	0,5871	0,4485
Accuracy Anziani	–	0,1053	0,1088
Lift	1,1305	1,1073	0,8320

Tabella 7.11: *confronto tra i migliori risultati ottenuti nell'inferenza dell'età degli utenti, utilizzando i classificatori di base separatamente e combinati tra loro mediante le due soluzioni proposte.*

Inferenza dell'età

Il confronto tra i migliori risultati ottenuti relativamente all'inferenza dell'età degli utenti è contenuta nella Tabella 7.11. In questo caso la metodologia basata sul Grading fornisce performance simili a quelle ottenute con SVM, mentre la seconda metodologia ottiene risultati decisamente inferiori in termini di Lift e Recall globale. La soluzione basata sull'uso di grader riesce ad assegnare utenti a tutte e quattro le classi previste, anche se i risultati in termini di Recall non sono particolarmente elevati.

La Tabella 7.12, invece, mostra le configurazioni che consentono di ottenere i risultati maggiormente equilibrati tra le varie classi. In questo caso le due metodologie di combinazione forniscono dei risultati migliori in termini di Lift, anche se entrambe sono inferiori ad uno.

In conclusione, è possibile notare come la metodologia basata sull'uso di Grading e Voting offra migliori prestazioni rispetto alla soluzione che utilizza l'Arbitro nella prima fase di classificazione. Relativamente agli obiettivi che ci siamo posti all'inizio del capitolo, invece, si può affermare che, se da un lato il ricorso alle soluzioni combinate consente una migliore distribuzione tra i risultati relativi alle varie classi, a livello globale si verifica una diminuzione

METRICA	Algoritmi di base (SVM)	Grading e Voting	Arbitro e Voting
Recall Globale	0,4034	0,4752	0,4393
Recall GIOVANISSIMI	0,3731	0,3731	0,1764
Recall GIOVANI	0,5293	0,5240	0,5768
Recall ADULTI	0,2605	0,4253	0,2866
Recall ANZIANI	0,2160	0,4000	0,3273
Accuracy Globale	0,4034	0,4752	0,4393
Accuracy GIOVANISSIMI	0,0868	0,1101	0,0916
Accuracy GIOVANI	0,5494	0,6895	0,6057
Accuracy ADULTI	0,3523	0,4745	0,4444
Accuracy ANZIANI	0,2411	0,1739	0,1088
Lift	0,7646	0,8795	0,8257

Tabella 7.12: *confronto tra i risultati più equilibrati ottenuti nell'inferenza dell'età degli utenti, utilizzando i classificatori di base separatamente e combinati tra loro mediante le due soluzioni proposte.*

delle performance, che non consente di esprimere un giudizio completamente positivo su tali tecniche. Le soluzioni da noi implementate, tuttavia, rappresentano solo un primo approccio all'argomento e potrebbero essere ulteriormente approfondite nell'ambito di ricerche future, ad esempio applicando differenti classificatori di base o sviluppando alcune delle altre metodologie presentate nel Paragrafo 2.4.

Capitolo 8

Porting

Indice

In questo capitolo viene presentata una tecnica, denominata Porting, utile per effettuare la caratterizzazione degli utenti in situazioni in cui non si abbiano a disposizione le informazioni demografiche necessarie alla costruzione del modello; dopo aver chiarito nel Paragarafo 8.1 i dettagli implementativi necessari per adoperare tale tecnica, nel Paragrafo 8.3 sono mostrati i risultati ottenuti attraverso la sua applicazione.

8.1 Introduzione

La tecnica presentata in questo lavoro prende il nome di *Porting*, e consente di ricavare un modello per la classificazione degli utenti da un dataset contenente sia informazioni demografiche sugli utenti che informazioni basate sui loro comportamenti e preferenze e di applicarlo per classificare utenti

appartenenti a dataset in cui non siano disponibili le informazioni anagrafiche indispensabili per ricavare un modello di classificazione analogo a quelli mostrati in precedenza.

La costruzione di un modello di classificazione realizzato tramite l'applicazione delle metodologie descritte nella prima parte del nostro lavoro necessita infatti di una notevole quantità di dati relativi agli utenti da profilare; alcune di queste informazioni, in particolare quelle anagrafiche, sono di difficile riperimento e non tutte le aziende interessate ne sono in possesso.

Inoltre, perchè i modelli costruiti applicando tali metodologie risultino efficaci, è necessario che le informazioni raccolte siano relative ad un numero significativo di utenti e coprano un periodo temporale sufficientemente esteso, in modo da valutare eventuali modifiche delle loro preferenze ed anticipare possibili trend futuri.

Queste difficoltà potrebbero quindi rappresentare una barriera per l'introduzione di tecniche basate sul Targeted Advertising, non essendoci le basi per usufruire delle funzionalità messe a disposizione dalle relative tecniche operative. In una situazione di questo tipo, il Porting si dimostra un'ottima soluzione anche da un punto di vista commerciale. Le aziende in possesso delle informazioni relative agli utenti possono infatti dar luogo a modelli per la classificazione, permettendo a chi ne ha bisogno di preoccuparsi soltanto della fase di profilazione, applicando le opportune metodologie in relazione al modello acquisito da una terza parte. É necessario in questo caso un riadattamento delle strutture dati da parte di chi utilzza il modello nei confronti di chi lo produce, operazione che non risulta essere particolarmente problematica e che può addirittura essere realizzata dal fornitore stesso del modello. In caso di utilizzo della matrice UGM, la questione si riduce alla riduzione o all'aumento del numero di colonne relative ai generi dei vari film, fondendo eventuali categorie simili in una sola di riferimento. Il problema è meno complesso di quanto possa sembrare, in quanto i vantaggi derivanti dalla possibilità di utilizzo delle tecniche di Targeted Advertising sono comunque superiori alla necessità di riadattamento delle strutture dati.

8.2 Schema implementativo

Nel seguito viene descritto lo schema implementativo seguito per l'applicazione della tecnica del Porting ai due dataset a nostra disposizione. Tale

procedura è stata eseguita sia utilizzando il dataset Yahoo! come insieme di apprendimento su cui costruire un modello da applicare al dataset MovieLens, sia invertendo i ruoli assegnati alle due basi di dati.

Per realizzare il Porting tra i due dataset sono richieste alcune modifiche agli schemi implementativi descritti in precedenza per l'applicazione delle metodologie tradizionali di classificazione, in particolare per quanto riguarda la fase di preparazione dei dataset: è stato infatti necessario effettuare un mapping dei generi presenti nelle due basi di dati da utilizzare, allo scopo di individuare quelli presenti in entrambi ed eliminare eventuali generi presenti in uno solo dei due dataset.

Attraverso una fase di analisi sono stati individuati 13 generi comuni, riportati nella Tabella 8.1. Per ottenere tali generi sono stati necessari alcuni accorgimenti: innanzi tutto i generi *Action* ed *Adventure* sono rappresentati separatamente nel dataset MovieLens, mentre si trovano uniti in Yahoo!; si è quindi provveduto ad aggregare i due generi anche nel dataset MovieLens, ottenendo il genere comune *Action/Adventure*. Inoltre il genere *Noir* presente soltanto in MovieLens è stato aggregato con il genere *Crime*, formando un unico genere *Crime/Noir*, comune ad entrambi i dataset.

Le matrici UGM relative ai due dataset sono quindi state modificate, eliminando le colonne relative ai generi da non considerare in quanto non comuni e sommando i valori contenuti nelle colonne relative ai generi che sono stati aggregati tra di loro. É stato quindi eseguito un ordinamento delle colonne, in modo che ciascuna di esse fosse associata con il medesimo genere di film in entrambi i dataset.

A partire dalle matrici così ricavate, sono state effettuate le operazioni di preparazione dei dataset: per le metodologie di analisi sono state svolte le medesime operazioni descritte nel Paragrafo 5.1.1, oltre all'eliminazione di eventuali record associati ad utenti i cui generi preferiti fossero stati tutti eliminati nella fase di mapping. I record relativi alle preferenze espresse da tali utenti, infatti, sarebbero stati formati da elementi tutti pari a zero, e avrebbero potuto generare errori e comporatamenti imprevisti per i vari algoritmi implementati.

Le successive fasi di costruzione del modello, ricerca dei migliori parametri ed esecuzione del test globale, sono state svolte seguendo le medesime modalità descritte nei capitoli precedenti.

Le Tabelle 8.1(a) e 8.1(b) riportano l'elenco dei generi originariamente

GENERE
Special Interest
Western
Action/Adventure
Suspence/Horror
Kids/Family
Comedy
Documentary
Thriller
Drama
Art/Foreign
Crime
Romance
Science-Fiction
Musical
Animation
Misc
Adult
Reality

(a) *Dataset Yahoo!.*

GENERE
Animation
Kids
Comedy
Adventure
Fantasy
Romance
Drama
Action
Crime
Thriller
Horror
Science-Fiction
Documentary
War
Musical
Mistery
Film Noir
Western

(b) *Dataset MovieLens.*

GENERE
Western
Action/Adventure
Suspence/Horror
Kids/Family
Comedy
Documentary
Thriller
Drama
Crime/Noir
Romance
Science-Fiction
Musical
Animation

(c) *Generi comuni.*

Tabella 8.1: *l'operazione di mapping dei generi dei film, effettuata per individuare i generi comuni tra i due dataset. In rosso sono evidenziati i generi di film che non sono stati considerati in quanto presenti in uno solo dei due dataset, in verde sono evidenziati i generi per cui sono state eseguite operazioni di aggregazione.*

presenti nei due dataset, la Tabella 8.1(c) contiene invece i 13 generi comuni individuati durante la fase di mapping. In rosso sono evidenziati i generi che non sono stati presi in considerazione, mentre i generi sui quali sono state compiute operazioni di aggregazione sono evidenziati in verde.

A titolo di esempio, nella Figura 8.1 è proposto lo schema implementativo per l'applicazione del Porting alla metodologia basata sulle Regole di Associazione. In questo caso viene utilizzato il dataset Yahoo! per apprendere il modello, il quale è poi applicato al dataset MovieLens per effettuare l'inferenza del Lifestyle degli utenti. Un procedimento analogo è stato seguito per il Porting da MovieLens a Yahoo! e per l'applicazione delle altre metodologie di data mining.

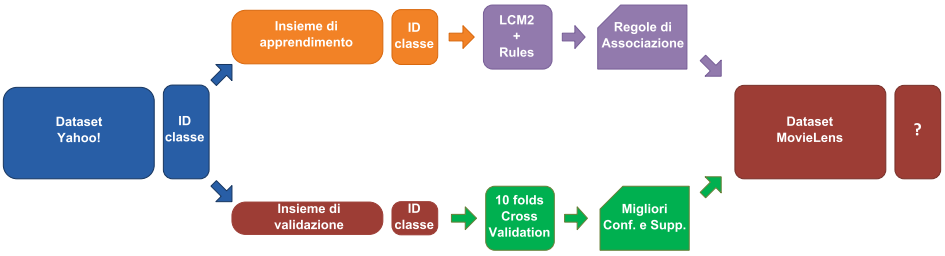


Figura 8.1: *Schema implementativo per il Porting dal dataset Yahoo! a MovieLens applicato alla metodologia di analisi basata sulle AR.*

8.3 Risultati ottenuti

In questo paragrafo vengono riportati solamente i risultati relativi ai test sull'inferenza del sesso degli utenti, in quanto non sono stati ottenuti risultati soddisfacenti per quanto riguarda la categorizzazione degli utenti in base all'età. Nell'ambito di tale tipologia di test, infatti, tutti gli utenti sono stati classificati come appartenenti alla classe maggiormente popolata, senza assegnare alcun utente alle classi minori; la Tabella 8.2 mostra un esempio dei risultati per questo tipo di test, ottenuti applicando SVD. In questo caso la Recall e la Accuracy complessiva coincidono con la percentuale di utenti appartenenti alla classe più popolata, mentre il Lift è pari ad uno, segno che la tecnica utilizzata non produce miglioramenti rispetto alle soluzioni tradizionali.

METRICA	ANZIANI	ADULTI	GIOVANI	GIOVANISSIMI	TOTALE
Recall	-	–	0,5294	-	-
Accuracy	-	-	0,5294	-	-
Utenti Assegnati	0	0	1871	0	0
Utenti Non Ass.	0	0	0	0	0
Recall Globale = 0,5294				Lift = 1,0000	

Tabella 8.2: *Porting - Esempio di risultati per l'inferenza dell'età, applicando la tecnica del Porting. Per questa tipologia di test i risultati ottenuti non possono essere considerati soddisfacenti.*

Effettuando un'analisi della composizione delle varie classi di età nei due dataset emerge come, in maniera opposta a quanto avviene per la suddivisione degli utenti in base al sesso, la ripartizione in base alla classe di età presenti notevoli differenze.

La Tabella 8.3 mostra, per ciascuna classe, il numero di utenti che la compongono oltre alla percentuale rispetto al numero complessivo di utenti nel dataset.

	Yahoo!	MovieLens
GIOVANISSIMI [<= 17]	2243 29,48%	225 3,68%
GIOVANI [>= 18 e < 35]	4165 54,78%	3197 52,94%
ADULTI [>= 35 e < 55]	1056 13,88%	2240 37,1%
ANZIANI [>= 56]	144 1,89%	378 6,28%

Tabella 8.3: *confronto tra la composizione delle classi di età nei dataset Yahoo! e MovieLens.*

La classe GIOVANI è l'unica per la quale la percentuale degli utenti rispetto al totale risulta simile in entrambi i dataset. Per le altre classi analizzate il numero di utenti e, di conseguenza, la percentuale rispetto al totale presentano delle differenze significative tra le due basi di dati, che rendono di fatto impossibile la costruzione di un modello comune da applicare nella classificazione degli utenti. La classe GIOVANISSIMI, ad esempio, è formata da circa il 30%

degli utenti nel dataset Yahoo!, mentre rappresenta solo il $3{,}68\%$ degli utenti del dataset MovieLens.

A partire da queste osservazioni è possibile affermare come l'applicazione della tecnica del Porting richieda un certo livello di correlazione tra la distribuzione degli utenti nei dataset utilizzati nelle fasi di apprendimento ed applicazione del modello. Tale necessità potrebbe risultare una limitazione all'applicazione del Porting, dato che il suo scopo principale consiste nell'operare senza conoscere le informazioni demografiche relative agli utenti su cui effettuare la classificazione. Una possibile soluzione potrebbe essere l'utilizzo nella fase di apprendimento di un dataset appositamente costruito, ottenuto ad esempio bilanciando gli utenti tra le varie classi, in modo da evitare che la presenza di classi nettamente più ampie rispetto alle altre possa determinare risultati poco soddisfacenti.

8.3.1 Porting da MovieLens a Yahoo!

In questa sezione sono presentati i risultati dei test in cui si è utilizzato il dataset MovieLens per la costruzione del modello per effettuare l'inferenza del sesso sugli utenti del dataset Yahoo!. Per quanto riguarda i test relativi all'applicazione del metodo basato sulle AR, il modello per la classificazione degli utenti è stato costruito selezionando casualmente $5\,000$ utenti dal dataset di apprendimento, a causa di problemi computazionali derivanti dall'utilizzo di un dataset troppo esteso; le dimensioni di tale campione sono comunque sufficienti per la costruzione di un modello affidabile da applicare nella fase di classificazione.

Regole di Associazione

La configurazione di parametri per la quale si sono ottenuti i migliori risultati relativamente all'applicazione del metodo basato sulle AR è mostrata nella Tabella 8.4.

In questo caso la maggioranza degli utenti viene assegnata alla classe MASCHI, per la quale si ricavano valori di Recall ed Accuracy pari rispettivamente a $0{,}9407$ e $0{,}7174$. Relativamente alla classe FEMMINE, invece, si ottengono una Recall del $6{,}4\%$ ed una Accuracy pari a $0{,}3309$, inferiori rispetto ai corrispettivi valori per la classe MASCHI, nonostante il numero di regole generato sia il medesimo per entrambe le classi.

	METRICA	MASCHI	FEMMINE	TOTALE
Best Recall	Recall	0,9407	0,0636	-
	Accuracy	0,7174	0,3309	-
	Utenti Assegnati	714	467	7609
	Utenti Non Assegnati	0	0	0
	Regole Generate	13	13	26
	Recall Globale = 0,6962		Lift = 0,9991	
	confidenza = 0,53		supporto = 5	

Tabella 8.4: *Porting da MovieLens a Yahoo! - La migliore configurazione di parametri per l'applicazione del metodo basato sulle AR per l'inferenza del sesso degli utenti.*

Singular Value Decomposition

A differenza di quanto avveniva nell'applicazione del metodo su un singolo dataset, nella configurazione in cui si utilizzano due colonne per la descrizione del Lifestyle si riescono ad assegnare utenti ad entrambe le classi previste. La Tabella 8.5 consente di confrontare i migliori risultati ottenuti per le due configurazioni.

	METRICA	MASCHI	FEMMINE	TOTALE
Best Recall 1 Colonna	Recall	0,9164	0,2036	-
	Accuracy	0,7418	0,4944	-
	Utenti Assegnati	6711	898	7609
	Utenti Non Assegnati	0	0	0
	Recall Globale = 0,7121		Lift = 0,9982	
Best Recall 2 Colonne	Recall	0,8454	0,3425	-
	Accuracy	0,7619	0,4710	-
	Utenti Assegnati	6023	1586	7609
	Utenti Non Assegnati	0	0	0
	Recall Globale = 0,7012		Lift = 0,9830	

Tabella 8.5: *Porting da MovieLens a Yahoo! - Confronto dei migliori risultati ottenuti utilizzando una e due colonne per la rappresentazione del sesso degli utenti, applicando il metodo basato su SVD.*

Confrontando i risultati ottenuti per le due modalità emerge come la soluzione basata sull'utilizzo di una sola colonna permetta di ottenere risultati

migliori a livello globale, ma con un minore bilanciamento tra i risultati relativi alle due classi. Nella prima soluzione, infatti, si ottiene una Recall pari a $0{,}9164$ per la classe MASCHI, mentre il corrispondente valore per la classe FEMMINE equivale a $0{,}2036$. La soluzione basata sull'uso di due colonne, invece, permette di ottenere una Recall inferiore per la classe MASCHI($0{,}8454$), alla quale corrisponde però una Recall relativa alla classe FEMMINE pari a $0{,}3425$.

Dal punto di vista dell'Accuracy è possibile notare come la prima soluzione sia globalmente più precisa rispetto alla seconda; relativamente alle singole classi, invece, la prima soluzione risulta più accurata nell'assegnazione degli utenti alla classe FEMMINE, per la quale è stato infatti registrato un valore inferiore di Recall, mentre tale soluzione origina un numero più elevato di errori nell'assegnazione degli utenti alla classe MASCHI.

Support Vector Machines

La migliore configurazione per questa tipologia di test è mostrata in Tabella 8.6, ed è ottenuta operando con SVM in modalità *Nu*-classification, applicando un kernel di tipo lineare e uno scaling compreso tra -1 ed 1 sulla matrice UGM non normalizzata. Per tale configurazione si ottengono una Recall globale pari a $0{,}7042$ ed un Lift inferiore ad uno.

	METRICA	MASCHI	FEMMINE	TOTALE
	Recall	$0{,}9569$	$0{,}0503$	-
	Accuracy	$0{,}7228$	$0{,}3107$	-
Best Recall	Recall Globale = $0{,}7042$		Lift = $0{,}9763$	
	C = $0{,}125$ n = $0{,}5$ UGM non normalizzata Kernel: LINEARE Scaling: $[-1, 1]$			

Tabella 8.6: *Porting da MovieLens a Yahoo! - La migliore configurazione di parametri per l'applicazione del metodo basato su SVM in modalita nu-classification per l'inferenza del sesso degli utenti.*

8.3.2 Porting da Yahoo! a MovieLens

Per effettuare i test di cui sono mostrati i risultati in questa sezione si è utilizzato il dataset Yahoo! per la costruzione del modello da utilizzare per la caratterizzazione degli utenti del dataset MovieLens.

Regole di Associazione

In questo tipo di test non sono stati ottenuti risultati soddisfacenti; per tutte le combinazioni di parametri utilizzate, infatti, tutti gli utenti vengono assegnati alla classe MASCHI, senza che siano prodotte regole per l'assegnazione di utenti alla classe FEMMINE. Il Lift ottenuto è quindi sempre uguale ad uno, mentre i valori di Accuracy e Recall globali sono equivalenti alla percentuale di utenti appartenenti alla classe più popolata (71,7%).

Singular Value Decomposition

La Tabella 8.7 mostra le configurazioni di parametri per cui si sono ottenuti i risultati migliori, utilizzando rispettivamente una e due colonne per descrivere il Lifestyle degli utenti.

	METRICA	MASCHI	FEMMINE	TOTALE
Best Recall 1 Colonna	Recall	0,9753	0,1492	-
	Accuracy	0,7434	0,7044	-
	Utenti Assegnati	5677	342	6039
	Utenti Non Assegnati	0	0	0
	Recall Globale = 0,7415		Lift = 1,0342	
Best Recall 2 Colonne	Recall	0,	0,	-
	Accuracy	0,7986	0,4807	-
	Utenti Assegnati	4275	1764	6039
	Utenti Non Assegnati	0	0	0
	Recall Globale = 0,70575		Lift = 0,9843	

Tabella 8.7: *Porting da Yahoo! a MovieLens - Confronto dei migliori risultati ottenuti utilizzando una e due colonne per la rappresentazione del sesso degli utenti, applicando il metodo basato su SVD.*

Per la prima configurazione, nella quale si utilizza una sola colonna per la descrizione del Lifestyle, si ottiene una Recall molto elevata per la classe

MASCHI, al quale corrisponde però, un valore pari a circa il 15% per la medesima metrica associata alla classe FEMMINE. I valori di Accuracy associati ad entrambe le classi, invece, sono superiori alla percentuale di utenti presenti; il risultato ottenuto per la classe FEMMINE, pari a $0{,}7044$, è molto superiore rispetto alla percentuale di utenti nella classe, pari a circa il 28%. Il Lift ottenuto per questa configurazione è pari a $1{,}0342$.

Utilizzando due colonne, invece, si ottengono prestazioni inferiori, testimoniate dal fatto che il Lift ottenuto sia inferiore all'unità.

Support Vector Machines

Come evidenziato dalla Tabella 8.8, in questo caso la migliore configurazione è ottenuta applicando la *C*-classification con kernel radiale ed uno scaling tra 0 e 1. L'unica affinità con la configurazione di parametri mostrata in Tabella 8.6, relativa alla migliore configurazione per i test da MovieLens a Yahoo!, è l'utilizzo di una matrice UGM non normalizzata, segnale che la procedura di normalizzazione potrebbe avere effetti negativi sull'applicazione di SVM nella metodologia del Porting.

	METRICA	MASCHI	FEMMINE	TOTALE
Best Recall	Recall	$0{,}9200$	$0{,}0823$	-
	Accuracy	$0{,}7255$	$0{,}2809$	-
	Recall Globale = $0{,}6897$		Lift = $0{,}9513$	
	C = 32768 $\gamma = 8$ UGM non normalizzata Kernel: RADIALE Scaling: [0, 1]			

Tabella 8.8: *Porting da Yahoo! a MovieLens - La migliore configurazione di parametri per l'applicazione del metodo basato su SVM in modalità C-classification per l'inferenza del sesso degli utenti.*

Al termine dell'analisi dei risultati ottenuti con l'applicazione della tecnica del Porting è possibile affermare come tale tecnica si dimostri utile per la costruzione di modelli da applicare a basi di dati che non forniscono informazioni demografiche sugli utenti. La necessità di una correlazione tra le basi di dati da utilizzare come insiemi di apprendimento e test potrebbe essere vista come una limitazione; tuttavia essa è superabile ottenendo le informazioni

demografiche relative ad un campione rappresentativo degli utenti da caratterizzare, in modo da individuare la distribuzione tra le varie classi ed applicare di conseguenza un modello ad essa correlato.

Capitolo 9

Conclusioni e sviluppi futuri

Al termine del nostro lavoro possiamo esprimere alcune considerazioni sui risultati ricavati e sui possibili scenari futuri legati allo studio da noi effettuato. Innanzitutto è necessario ribadire la validità di algoritmi come AR e SVD nell'ambito del Targeted Advertising: i test esaustivi effettuati sui parametri specifici di tali algoritmi, infatti, consentono di ottenere buoni risultati sia dal punto di vista della Recall, sia dal punto di vista della nuova metrica introdotta, il Lift. Quest'ultima si rivela uno strumento di grande potenzialità, dimostrando la sua efficacia nel valutare in maniera più approfondita l'impatto delle soluzioni adottate rispetto ai casi in cui si faccia affidamento a strategie di pubblicità tradizionali e fornendo al tempo stesso nuovi spunti di analisi rispetto alle tecniche tradizionali. Il fatto di aver ottenuto valori di Lift superiori ad uno per le metodologie basate su AR e SVD contribuisce a riaffermare l'utilità di tali algoritmi per l'inferenza del Lifestyle nell'ambito di soluzioni volte a proporre pubblicità mirata.

L'impiego delle Support Vector Machines propone la possibilità di utilizzo di una nuova metodologia come strumento di Targeted Advertising: la classificazione degli utenti si dimostra infatti accurata e precisa per entrambe le tipologie di test effettuate, fornendo ottime performance anche dal punto di vista computazionale e risultati generalmente migliori di quelli ottenuti con AR ed SVD, sia in termini di Recall globale che di Lift. In particolare per il dataset MovieLens sono stati ricavati valori di Lift pari a $1{,}0971$ e $1{,}1320$, rispettivamente per test basati sull'inferenza del sesso e dell'età mentre, per quanto riguarda le Recall globali, i migliori risultati equivalgono a $0{,}7925$ e $0{,}6071$. Da questo punto di vista il nostro lavoro lascia inoltre spazio ad un numero

significativo di possibili sviluppi futuri: se gli studi effettuati relativamente all'utilizzo di AR e SVD dovrebbero infatti avere consentito di completare l'analisi già iniziata nei lavori di tesi precedenti, per quanto riguarda le SVM le possibili sperimentazioni finalizzate all'incremento delle performance sono numerose e potrebbero riguardare, ad esempio, la scelta di nuovi tipi di kernel da utilizzare o l'applicazione di differenti configurazioni di parametri.

Relativamente alle basi di dati utilizzate, i risultati ottenuti con il nuovo dataset Yahoo! appaiono globalmente inferiori a quelli relativi al dataset MovieLens. La scarsa disponibilità di basi di dati liberamente accessibili su cui applicare le metodologie di analisi rappresenta una delle limitazioni principali nello sviluppo di soluzioni per il Targeted Advertising, per cui è auspicabile che, visto il notevole interesse dimostrato dal mercato verso questo ambito, si possa assistere alla diffusione di tali risorse.

La combinazione degli algoritmi è stata realizzata con l'obiettivo di effettuare classificazioni più accurate ed ottenere in questo modo una migliore e più corretta assegnazione degli utenti, anche in presenza di classi la cui dimensione risulti molto inferiore rispetto a quella di altre. La tecnica è stata implementata correttamente, consentendo di ottenere Lift superiori ad uno nella maggior parte delle tipologie di test effettuati; ad ogni modo, i risultati raggiunti sono complessivamente inferiori rispetto a quelli ottenuti dall'applicazione di SVM. A partire da tale considerazione è possibile delineare le caratteristiche di eventuali studi futuri destinati ad un incremento delle performance in tale ambito: relativamente alle due metodologie proposte, è infatti possibile pensare di valutare soglie alternative e di assegnare nuovi pesi legati all'output dei vari algoritmi, nel tentativo di ottenere configurazioni più performanti rispetto a quelle sperimentate. Un altro aspetto che potrebbe risultare interessante è dato dalla possibilità di sperimentare nuovi classificatori di base da utilizzare nelle prime fasi di tali soluzioni. Ulteriori approfondimenti potrebbero infine derivare dall'implementazione di metodologie alternative, basate su tecniche di combinazione differenti rispetto a quelle da noi proposte.

Infine, le sperimentazioni effettuate sulla tecnica del Porting forniscono risultati incoraggianti per la sua utilità nelle soluzioni di Targeted Advertising. Come evidenziato dall'analisi dei risultati, tuttavia, tale tecnica fornisce le migliori performance quando si verifica una correlazione tra le due basi di dati utilizzate nel procedimento. Ciò comporterebbe la necessità di ottenere le informazioni demografiche relative ad un campione rappresentativo degli

utenti da caratterizzare, in modo da individuare la distribuzione tra le varie classi ed applicare di conseguenza un modello ad essa correlato. Dal punto di vista applicativo, in ogni caso, si ricava comunque un beneficio dall'applicazione del Porting, dato che il quantitativo di dati necessari alla costruzione del campione è decisamente inferiore rispetto a quello che servirebbe per realizzare un modello sufficientemente accurato. Gli sviluppi futuri legati alla sperimentazione del Porting potrebbero riguardare l'utilizzo di un insieme di apprendimento maggiormente bilanciato, in modo da favorire la classificazione degli utenti delle classi più piccole, oltre all'applicazione di metodologie di combinazione di classificatori nella fase di apprendimento.

Appendice A

Risultati per AR ed SVD al variare del numero di generi

Nel seguito sono contenuti i risultati dei test realizzati applicando le metodologie basate su AR e SVD e variando il numero di generi preferiti tenuti in considerazione; in particolare sono stati effettuati test considerando uno, due e tre generi per ciascun utente. Come evidenziato nei capitoli precedenti, le migliori prestazioni per questo tipo di test si ricavano scegliendo di considerare tre generi di film. Per entrambi i dataset utilizzati sono proposti i risultati dei test sull'inferenza del sesso e dell'età degli utenti; relativamente a ciascuna tipologia sono presentate le configurazioni che consentono di ottenere le migliori performance in termini di Lift e di Recall complessiva. Per ciascuna di esse sono mostrati in particolare:

- il numero di generi considerati;
- le combinazioni di parametri applicate (confidenza e supporto per AR, k e t per SVD);
- i risultati ottenuti per ciascuna delle metriche di valutazione globali considerate (Lift, Recall ed Accuracy);
- il numero di classi alle quali sono stati assegnati utenti.

A.1 Dataset MovieLens

A.1.1 Inferenza del sesso

Regole di Associazione

Applicando le AR nei test per l'inferenza del sesso degli utenti appartenenti a MovieLens, la configurazione in cui si considerano tre generi di film garantisce i risultati migliori in termini di Lift e Recall. Come mostrato dalla Tabella A.1, utilizzando un solo genere di film si ottiene un valore di Lift più elevato, ma si assegnano utenti solo alla classe MASCHI.

	Generi	Conf.	Supp.	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	0,49	5	1,0290	0,7360	0,7360	2 su 2
	2	0,55	5	1,0351	0,7421	0,74214	2 su 2
	3	0,51	35	1,0440	0,7488	0,7488	2 su 2
Best Lift	1	0,01	95	1,0548	0,6476	0,7632	1 su 2
	2	0,55	5	1,0351	0,7421	0,7421	2 su 2
	3	0,51	35	1,0440	0,7488	0,7488	2 su 2

Tabella A.1: *dataset MovieLens - Risultati per l'inferenza del sesso degli utenti applicando le AR, in base al numero di generi di film considerati.*

Singular Value Decomposition

Nei test per l'inferenza del sesso degli utenti di MovieLens in cui si utilizza SVD i migliori risultati si ottengono considerando tre generi di film, sia per quanto riguarda la configurazione in cui il Lifestyle degli utenti viene rappresentato con una colonna A.2, sia nel caso in cui se ne utilizzino due A.3.

1 Colonna							
	Generi	$best_k$	$best_t$	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	1	0,01	1,0361	0,7445	0,7445	2 su 2
	2	9	−0,06	1,0361	0,7445	0,7445	2 su 2
	3	7	−0,12	1,0484	0,7533	0,7533	2 su 2
Best Lift	1	1	0,01	1,0361	0,7445	0,7445	2 su 2
	2	9	−0,06	1,0361	0,7445	0,7445	2 su 2
	3	7	−0,12	1,0484	0,7533	0,7533	2 su 2

Tabella A.2: *dataset MovieLens - Risultati per l'inferenza del sesso degli utenti applicando SVD con una colonna per la descrizione del Lifestyle, in base al numero di generi di film considerati.*

2 Colonne							
	Generi	$best_k$	$best_t$	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	1	−0,5	1,0000	0,7185	0,7185	1 su 2
	2	8	0	1,0054	0,7224	0,7224	2 su 2
	3	7	−0,5	1,0261	0,7373	0,7373	2 su 2
Best Lift	1	1	−0,5	1,0000	0,7185	0,7185	1 su 2
	2	8	0	1,0054	0,7224	0,7224	2 su 2
	3	7	−0,5	1,0261	0,7373	0,7373	2 su 2

Tabella A.3: *dataset MovieLens - Risultati per l'inferenza del sesso degli utenti applicando SVD con due colonne per la descrizione del Lifestyle, in base al numero di generi di film considerati.*

A.1.2 Inferenza dell'età

Regole di Associazione

La Tabella A.4 mostra i risultati ottenuti dall'applicazione del metodo basato sulle AR per l'inferenza dell'età degli utenti del dataset MovieLens. Si può notare come, nel caso della configurazione che assicuri la migliore Recall globale, i risultati ottenuti applicando due o tre generi siamo gli stessi. Anche dal punto di vista del miglior Lift per tali scelte di generi le performance sono molto simili.

	Generi	Conf.	Supp.	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	0,45	5	1,0658	0,5561	0,5561	2 su 4
	2	0,43	35	1,0710	0,5672	0,5672	2 su 4
	3	0,43	35	1,0710	0,5672	0,5672	2 su 4
Best Lift	1	0,53	5	1,0741	0,3346	0,6001	2 su 4
	2	0,59	35	1,0844	0,4288	0,5998	2 su 4
	3	0,59	35	1,0862	0,4415	0,5896	2 su 4

Tabella A.4: *dataset MovieLens - Risultati per l'inferenza dell'età degli utenti applicando le AR, in base al numero di generi di film considerati.*

Singular Value Decomposition

Per quanto riguarda l'applicazione di SVD per caratterizzare gli utenti di MovieLens in base all'età, il miglior Lift si ottiene considerando un solo genere (Tabella A.5); anche in questo caso, tuttavia, tale scelta consente di assegnare utenti ad una sola classe. Questa tipologia di test è l'unica per il dataset MovieLens in cui le performance migliori non si ottengono considerando tre generi: infatti la migliore Recall globale si ottiene mantenendo solamente due generi per ciascun utente.

	Generi	$best_k$	$best_t$	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	3	$-0{,}5$	0,9753	0,5237	0,5237	2 su 4
	2	5	$-0{,}5$	1,0401	0,5585	0,5585	2 su 4
	3	6	$-0{,}5$	1,0092	0,5419	0,5419	2 su 4
Best Lift	1	1	0,14	1,0659	0,4194	0,5864	1 su 4
	2	6	0,47	1,0088	0,5408	0,5408	2 su 4
	3	8	0,5	1,0515	0,4796	0,5698	2 su 4

Tabella A.5: *dataset MovieLens - Risultati per l'inferenza dell'età degli utenti applicando SVD con una colonna per la descrizione del Lifestyle, in base al numero di generi di film considerati.*

A.2 Dataset Yahoo!

A.2.1 Inferenza del sesso

Regole di Associazione

La Tabella A.6 mostra i migliori risultati per i test sull'inferenza del sesso degli utenti del dataset Yahoo!, basati su AR. Anche in questo caso mantenendo un solo genere di film si ricava un Lift migliore che considerandone tre, ma la prima soluzione non assegna alcun utente alla classe FEMMINE.

	Generi	Conf.	Supp.	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	0,49	5	1,0120	0,7215	0,7223	2 su 2
	2	0,53	35	1,0168	0,7253	0,7255	2 su 2
	3	0,51	35	1,0271	0,7328	0,7328	2 su 2
Best Focused	1	0,59	185	1,0505	0,5275	0,7658	1 su 2
	2	0,53	35	1,0168	0,7253	0,7255	2 su 2
	3	0,51	35	1,0271	0,7328	0,7328	2 su 2

Tabella A.6: *dataset Yahoo! - Risultati per l'inferenza del sesso degli utenti applicando le AR, in base al numero di generi di film considerati.*

Singular Value Decomposition

Per quanto riguarda questa tipologia di test i migliori risultati si ottengono utilizzando una colonna per rappresentare il Lifestyle dell'utente (Tabella A.7). La configurazione che impiega due colonne a tale scopo (Tabella A.8), invece, assegna tutti gli utenti alla classe MASCHI; di conseguenza il Lift è uguale ad uno, mentre l'Accuracy e la Recall globale sono pari alla percentuale di utenti appartenenti a tale classe nel campione selezionato.

1 Colonna							
	Generi	Conf.	Supp.	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	8	−0,08	1,0079	0,7231	0,7231	2 su 2
	2	2	−0,33	1,0043	0,7201	0,7201	2 su 2
	3	2	−0,5	1,0188	0,7309	0,7309	2 su 2
Best Lift	1	9	−0,07	1,0140	0,7270	0,7270	2 su 2
	2	2	−0,33	1,0043	0,7201	0,7201	2 su 2
	3	2	−0,48	1,0201	0,7314	0,7314	2 su 2

Tabella A.7: *dataset Yahoo! - Risultati per l'inferenza del sesso degli utenti applicando SVD con una colonna per la descrizione del Lifestyle, in base al numero di generi di film considerati.*

SVD - 2 Colonne							
	Generi	Conf.	Supp.	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	1	−0,5	1,0000	0,7159	0,7159	1 su 2
	2	1	−0,5	1,0000	0,7159	0,7159	1 su 2
	3	1	−0,5	1,0000	0,7159	0,7159	1 su 2
Best Lift	1	1	−0,1	1,0000	0,7159	0,7159	1 su 2
	2	1	−0,5	1,0000	0,7159	0,7159	1 su 2
	3	1	−0,5	1,0000	0,7159	0,7159	1 su 2

Tabella A.8: *dataset Yahoo! - Risultati per l'inferenza del sesso degli utenti applicando SVD con due colonne per la descrizione del Lifestyle, in base al numero di generi di film considerati.*

A.2.2 Inferenza dell'età

Regole di Associazione

La Tabella A.9 mostra i migliori risultati per i test per la caratterizzazione dell'età degli utenti del dataset Yahoo!, basati su AR.

	Generi	Conf.	Supp.	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	0,49	5	1,0270	0,5326	0,5326	2 su 2
	2	0,01	35	1,0400	0,5336	0,5336	2 su 2
	3	0,51	35	1,0506	0,5654	0,5654	2 su 2
Best Lift	1	0,53	5	1,0215	0,5142	0,5142	2 su 2
	2	0,59	155	1,0249	0,3726	0,6139	2 su 2
	3	0,59	185	1,0325	0,4459	0,5964	2 su 2

Tabella A.9: *dataset Yahoo! - Risultati per l'inferenza dell'età degli utenti applicando le AR, in base al numero di generi di film considerati.*

Singular Value Decomposition

I risultati relativi ai test di inferenza dell'età per il dataset Yahoo! sono riportati nella Tabella A.10.

	Generi	$best_k$	$best_t$	Lift	Recall globale	Accuracy globale	Classi Ass.
Best Recall	1	2	$-0,5$	1,0170	0,5275	0,5275	2 su 2
	2	4	$-0,2$	0,9969	0,5161	0,5161	2 su 2
	3	6	0,5	1,0212	0,5473	0,5473	2 su 2
Best Lift	1	2	0,21	1,0053	0,5114	0,6017	2 su 2
	2	1	$-0,48$	1,0149	0,5149	0,5898	2 su 2
	3	1	0,5	1,0025	0,5084	0,6214	2 su 2

Tabella A.10: *dataset Yahoo - Risultati per l'inferenza dell'età degli utenti applicando SVD con una colonna per la descrizione del Lifestyle, in base al numero di generi di film considerati.*

Bibliografia

[1] R. Agrawal, T. Imieliński, and A. Swami. Mining association rules between sets of items in large databases. *SIGMOD Rec.*, 22:207–216, June 1993.

[2] S. M. Bae, S. H. Ha, and S. C. Park. Fuzzy web ad selector based on web usage mining. *IEEE Intelligent Systems*, 18:62–69, November 2003.

[3] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1st edition, May 1999.

[4] E. C. Baird. Targeted online advertising: persuasion in an era of massless communication. Master's thesis, Massachussets Institute Of Technology, 2008.

[5] C. Becquet, S. Blachon, B. Jeudy, J.F. Boulicaut, and O. Gandrillon. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. *Genome Biology*, 3(12):1–16, 2002.

[6] M. W. Berry. Large scale sparse singular value computations. *International Journal of Supercomputer Applications*, 6:13–49, 1992.

[7] L. Bianchi. Algoritmi per la selezione di pubblicità mirata nella televisione interattiva. Master's thesis, Politecnico di Milano, 2008.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

[9] T. Bozios, G. Lekakos, and V. Skoularidou. Advanced techniques for personalized advertising in a digital tv environment: The imedia system. In *In Proceedings of the eBusiness and eWork Conference*, 2001.

[10] P. K. Chan and S. J. Stolfo. Experiments on multistrategy learning by meta-learning. In *Proceedings of the second international conference on Information and knowledge management*, CIKM '93, pages 314–323, New York, NY, USA, 1993. ACM.

[11] P. K. Chan and S. J. Stolfo. Learning arbiter and combiner trees from partitioned data for scaling machine learning. In AAAI Press, editor, *In Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, pages 39–44, 1995.

[12] C. Cortes, S. Kumar, A. Makadia, G. Mann, J. Yagnik, and M. Zhao. Video content analysis for automatic demographics recognition of users and videos, 2010.

[13] C. Cortes and V. Vapnik. Support-vector networks. *Mach. Learn.*, 20:273–297, September 1995.

[14] B. V. Dasarathy and B. V. Sheela. A composite classifier system design: concepts and methodology. In *Proceedings of IEEE*, page 67:708Ű713, 1978.

[15] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1:211–218, 1936. 10.1007/BF02288367.

[16] S. Encheva and S. Tumin. Application of association rules in education. In De-Shuang Huang, Kang Li, and George Irwin, editors, *Intelligent Control and Automation*, volume 344 of *Lecture Notes in Control and Information Sciences*, pages 834–838. Springer Berlin / Heidelberg, 2006.

[17] A. Galeotti and J. L. Moraga-González. A model of strategic targeted advertising. *CESifo Working Paper Series*, 2004.

[18] G. Giacinto and F. Roli. Design of effective neural network ensembles for image classification processes. *Image Vision and Computing Journal*, 19(9-10):699–707, 2001.

[19] F. Granara. Categorizzazione di utenti iptv. Master's thesis, Politecnico di Milano, 2010.

[20] B. Gunter and A. Furnham. *Consumer profiles: an introduction to psychographics.* London: Routledge., 1992.

[21] I. Hawkins, K. A. Coney, and R. J. Best. Consumer behavior: Building marketing strategy. *Applied Mathematics and Computation*, 1998.

[22] T. K. Ho. Multiple classifier combination: Lessons and the next steps. *Hybrid Methods in Pattern Recognition*, 1:171–198, 2002.

[23] J. Hu, H. J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 151–160, New York, NY, USA, 2007. ACM.

[24] Y.S. Huang and C.Y. Suen. The behavior-knowledge space method for combination of multiple classifiers. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 347 – 352, 1993.

[25] J. Jaworska and M. Sydow. Behavioural targeting in on-line advertising: An empirical study. In *Proceedings of the 9th international conference on Web Information Systems Engineering*, WISE '08, pages 62–76, Berlin, Heidelberg, 2008. Springer-Verlag.

[26] J. W. Kim, B. H. Lee, M. J. Shaw, H. Chang, and M. Nelson. Application of decision-tree induction techniques to personalized advertisements on internet storefronts. *Int. J. Electron. Commerce*, 5:45–62, March 2001.

[27] L. I. Kuncheva. Clustering-and-selection model for classifier combination. In *Knowledge-Based Intelligent Engineering Systems and Allied Technologies. Proceedings. Fourth International Conference.*, volume 1, pages 185 – 188, 2000.

[28] L. I. Kuncheva. Switching between selection and fusion in combining classifiers: an experiment. *IEEE Trans Syst Man Cybern B Cybern*, 32(2):146–56, 2002.

[29] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[30] George Lekakos. Personalized advertising services through hybrid recommendation methods: the case of digital interactive television.

[31] J.i Liu, C. Wang, Z. Liu, and W. Yao. Advertising keywords extraction from web pages. In *Proceedings of the 2010 international conference on Web information systems and mining*, WISM'10, pages 336–343, Berlin, Heidelberg, 2010. Springer-Verlag.

[32] R. Liu and B. Yuan. Multiple classifier combination by clustering and selection. *Information Fusion*, 2:163–168, 2001.

[33] A. Metwally, D. Agrawal, and A. E. Abbadi. Using association rules for fraud detection in web advertising networks. In *Proceedings of the 31st international conference on Very large data bases*, VLDB '05, pages 169–180. VLDB Endowment, 2005.

[34] Nielsen preView. *Beyond the Ratings... How to buy/sell cinema based on robust consumer segmentations*.

[35] M. B. Oliver, S. L. Sargent, and J. B. Weaver. The Impact of Sex and Gender Role Self-Perception on Affective Reactions to Different Types of Film. *Sex Roles*, 38:45–62, 1998.

[36] J. Richards and D. Sheridan. *Mass-Observation at the movies*. Cinema and society. Routledge & Kegan, 1987.

[37] B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In *Proceedings, First International Conference on Knowledge Discovery & Data Mining, Menlo Park*, pages 252–257. AAAI Press, 1995.

[38] B. Schölkopf, K. Sung, C. Burges, F. Girosi, P. Niyogi, T. Poggio, and V. Vapnik. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE TRANSACTIONS ON SIGNAL PROCESSING*, 45:2758–2765, 1997.

[39] A. K. Seewald and J. Fürnkranz. An evaluation of grading classifiers. In *Proceedings of the 4th International Conference on Advances in Intelligent Data Analysis*, IDA '01, pages 115–124, London, UK, UK, 2001. Springer-Verlag.

[40] M. Senzani. Tecniche per la valutazione di algoritmi di raccomandazione. Master's thesis, Politecnico di Milano, 2007.

[41] W. E. Spangler, M. Gal-Or, and J. H. May. Using data mining to profile tv viewers. *Commun. ACM*, 46:66–72, December 2003.

[42] R. O. N. Tamborini and J. Stiff. Predictors of Horror Film Attendance and Appeal: An Analysis of the Audience for Frightening Films. *Communication Research*, 14(4):415–436, August 1987.

[43] T. Uno, M. Kiyomi, and H. Arimura. Lcm ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. In *FIMI'04*, pages –1–1, 2004.

[44] V. Vapnik. *Estimation of Dependences Based on Empirical Data (Information Science and Statistics)*. Springer, March 2006.

[45] V. Vapnik and A. Chervonenkis. A note on one class of perceptrons. *Automation and Remote Control*, 25, 1964.

[46] V. Vapnik and A. Lerner. Pattern recognition using generalized portrait method. *Automation and Remote Control*, 24, 1963.

[47] V. N. Vapnik and A. Ya. Chervonenkis. *Theory of Pattern Recognition*. Nauka, USSR, 1974.

[48] X. J. Wang, M. Yu, L. Zhang, R. Cai, and W. Y. Ma. Argo: intelligent advertising by mining a user's interest from his photo collections. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, ADKDD '09, pages 18–26, New York, NY, USA, 2009. ACM.

[49] X. J. Wang, L. Zhang, X. Li, and W. Y. Ma. Annotating images by mining image search results. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30:1919–1932, 2008.

[50] Sito web di Auditel. http://www.auditel.it.

[51] Sito web di Claritas PRIZME. http://www.claritas.com/claritas.

[52] Sito web di dmoz. http://www.dmoz.org.

[53] Sito web di Eurisko. http://www.gfk.com/gfk-eurisko.

[54] Sito web di GroupLens. http://www.grouplens.org.

[55] Sito web di lcm. http://research.nii.ac.jp/ uno/code/lcm.html.

[56] Sito web di Nielsen. http://www.nielsenmedia.com.

[57] Sito web di Rules. http://adrem.ua.ac.be/ goethals/software.

[58] Sito web di Yahoo Research. http://research.yahoo.com.

[59] Sito web flickr. http://www.flickr.com.

[60] Sito web si Sric. http://www.sric-bi.com/VALS.

[61] Sito web youtube. http://www.youtube.com.

[62] D. H. Wolpert. Original contribution: Stacked generalization. *Neural Netw.*, 5:241–259, February 1992.

[63] K. Woods, W. P. Kegelmeyer, Jr., and K. Bowyer. Combination of multiple classifiers using local accuracy estimates. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:405–410, 1997.

[64] J. Yan, N. Liu, G. Wang, W. Zhang, Y. Jiang, and Z. Chen. How much can behavioral targeting help online advertising? In *World Wide Web Conference Series*, pages 261–270, 2009.

[65] W. S. Yang, J. B. Dia, H. C. Cheng, and H. T. Lin. Mining social networks for targeted advertising. *Hawaii International Conference on System Sciences*, 6:137a, 2006.