Università di Pisa	A.A. 2006- 2007
Data Mining	
Corso di Laurea Specialistica in Informatica per l'economia e l'Azienda	

Progetti

Informazioni generali

Di seguito vengono proposti 7 progetti, ognuno rivolto a un gruppo di 3 persone. Per tutti viene richiesto di:

1. svolgere il lavoro di analisi indicato, seguendo i passi generali del modello CRISP, dalla comprensione del problema fino alla validazione dei risultati ottenuti;
2. produrre una relazione scritta che descrive in modo succinto le fasi svolte nel corso del proprio lavoro, seguendo anche qui (seppur con un certo grado di flessibilità) i passi indicati dal modello CRISP. Indicativamente la relazione dovrà occupare tra le 20 e le 30 pagine;
3. preparare una presentazione di circa 30 minuti, da effettuare in sede di esame da tutti i membri del gruppo.

Viene lasciata piena libertà in quanto agli strumenti utilizzati nello svolgimento dei progetti, purché tutte le fasi da essi richieste (esplorazione dei dati, preprocessing, mining, validazione, ecc.) siano adeguatamente coperte.

La persona di riferimento comune ad ogni progetto, cui rivolgersi per chiarimenti e problemi vari è Mirco Nanni (mirco.nanni@isti.cnr.it), affiancato in diversi progetti da un esperto (o quasi) del dominio applicativo.

Progetto 1A (COOP-defezione A)

Nell'ambito dell'Unicoop Tirreno vengono forniti i dati che descrivono gli acquisti dei clienti relativi ad un punto vendita, per un periodo di circa 8 mesi. Tali dati contengono il dettaglio dei singoli scontrini emessi dal punto vendita, ovvero ogni singolo prodotto venduto, nonché, per i clienti che sono soci della cooperativa, il codice che identifica univocamente il socio in questione. Vengono forniti inoltre: (a) le tabelle che descrivono la gerarchia di prodotti presenti nel circuito Unicoop, e (b) l'anagrafica dei soci.

Obiettivi del progetto:

1. dividere il dataset in due insiemi: *presente* e *futuro*. Il primo copre i primi 5 mesi di vendite, il secondo la rimanente parte;
2. in base al dataset *presente* caratterizzare ogni socio tramite una serie di attributi significativi, quali: dati anagrafici, comportamenti di acquisto (es.: quali e quanti prodotti o categorie di prodotti acquistano) e loro evoluzione nel tempo (es.: aumento o diminuzione degli acquisti, anche in settori specifici di prodotti);
3. in base al dataset *futuro* (ed anche a quello *presente*, se necessario), definire almeno 2 nozioni/definizioni diverse di "defezione" dei clienti da implementare in un omonimo attributo target (defezione1, defezione2, ...), ovvero: per ogni socio, determinare se nel periodo futuro ha abbandonato o sta probabilmente abbandonando il punto vendita Coop analizzato. Per ogni nozione di defezione scelta, analizzare la distribuzione dei defezionanti, e verificare in quale misura le diverse definizioni producono risultati diversi; (suggerimento: alcune delle possibili *grandezze* da considerare: il volume di spesa, la frequenza di acquisto, l'acquisto di prodotti di determinate categorie (es.: prodotti freschi, pesce,..., ecc.)
4. costruire uno o più classificatori in grado di prevedere la defezione dei soci secondo le definizioni scelte, ovvero, in grado di stabilire il valore corretto della variabile defezione (estratta dal dataset *futuro*) a partire dalle variabili selezionate al punto 2 (estratte dal dataset *presente*). Fornire una validazione dei classificatori prodotti e, analogamente a quanto fatto al punto 3, confrontare le predizioni ottenute per le diverse definizioni onde valutare quanto fortemente sono correlare.

Progetto 1B (COOP-defezione B)

Nell'ambito dell'Unicoop Tirreno vengono forniti i dati che descrivono gli acquisti dei clienti relativi ad un punto vendita, per un periodo di circa 8 mesi. Tali dati contengono il dettaglio dei singoli scontrini emessi dal punto vendita, ovvero ogni singolo prodotto venduto, nonché, per i clienti che sono soci della cooperativa, il codice che identifica univocamente il socio in questione. Vengono forniti inoltre: (a) le tabelle che descrivono la gerarchia di prodotti presenti nel circuito Unicoop, e (b) l'anagrafica dei soci.

Obiettivi del progetto:

5. dividere il dataset in due insiemi: *presente* e *futuro*. Il primo copre i primi 5 mesi di vendite, il secondo la rimanente parte;
6. in base al dataset *presente* caratterizzare ogni socio tramite una serie di attributi significativi, quali: dati anagrafici, comportamenti di acquisto (es.: quali e quanti prodotti o categorie di prodotti acquistano) e loro evoluzione nel tempo (es.: aumento o diminuzione degli acquisti, anche in settori specifici di prodotti);
7. in base al dataset *futuro* (ed anche a quello *presente*, se necessario), definire una o piu' nozioni di "defezione" dei clienti da implementare in un omonimo attributo target, ovvero: per ogni socio, determinare se nel periodo futuro ha abbandonato o

sta probabilmente abbandonando il punto vendita Coop analizzato;

8. trovare (almeno) una segmentazione dei soci rispetto alle variabili selezionate al punto 2 e cercare di fornire un profilo di ogni cluster individuato, anche in riferimento alla variabile "defezione";
9. costruire uno o più classificatori in grado di prevedere la defezione dei soci, ovvero, in grado di stabilire il valore corretto della variabile defezione (estratta dal dataset *futuro*) a partire dalle variabili selezionate al punto 3 (estratte dal dataset *presente*). Fornire anche una validazione dei classificatori prodotti.

Progetto 2 (COOP-market basket analysis)

A partire dagli stessi dati descritti nel progetto 1, si stabiliscono i seguenti obiettivi:

1. sull'intero periodo di tempo monitorato, caratterizzare ogni socio tramite una serie di attributi significativi, quali: dati anagrafici, comportamenti di acquisto (es.: quali e quanti prodotti o categorie di prodotti acquistano);
2. trovare una segmentazione dei soci rispetto alle variabili selezionate al punto 1 e cercare di fornire un profilo di ogni cluster individuato. Scegliere o cercare un numero adeguato di cluster, in particolare un numero non troppo elevato, dato che le analisi che seguono andranno eseguite separatamente su ogni cluster;
3. per ogni segmento di soci trovato al punto precedente, estrarre un insieme di regole associative, possibilmente mantenendo tra i diversi cluster valori non troppo diversi dei parametri (mimimo supporto e confidenza). Nota: occorre scegliere un adeguato livello di "astrazione" dei prodotti, in quanto trovare regole associative sui singoli prodotti (prodotto X di marca Y confezione Z...) è in genere impossibile;
4. fornire un profilo di ogni cluster in base alle regole associative che lo caratterizzano e distinguono dagli altri.

Progetto 3 (COOP-circolarità)

Uno dei fenomenti di interesse nella grande distribuzione è la cosiddetta "circolarità", ovvero il fatto che un socio possa frequentare diversi punti vendita della stessa catena con intensità diverse e anche variabili nel tempo. Ad esempio, alcuni punti vendita in centro città possono essere privilegiati da alcuni soci durante l'inverno, mentre in estate per gli stessi soci prevalgono o punti vendita più prossimi alla riviera.

Nell'ambito dell'Unicoop Tirreno vengono forniti i dati che descrivono gli acquisti effettuati in un certo numero di punti vendita (tra i 10 e i 20) tra loro relativamente vicini, per un periodo di circa un anno. Gli acquisti sono descritti in forma aggregata per ogni scontrino, ovvero non viene fornito il contenuto dettagliato di ogni scontrino o carrello., e sono collegabili all'anagrafica dei soci, anche essa fornita. Infine, viene data

anche una (seppur scarna) anagrafica dei negozi.
Gli obiettivi del progetto sono i seguenti:

1. caratterizzare ogni socio tramite una serie di attributi significativi, quali: dati anagrafici, comportamenti di acquisto (es.: quali e quanti prodotti o categorie di prodotti acquistano);
2. trovare una segmentazione dei soci rispetto alle variabili selezionate al punto 1 e cercare di fornire un profilo di ogni cluster individuato. Inoltre, analizzare la distribuzione dei diversi cluster all'interno di ogni punto vendita. Nota: proprio per la circolarità menzionata all'inizio, un socio può appartenere a più punti vendita. Trattare questi casi nel modo ritenuto più opportuno, giustificando la scelta;
3. si definisca *gruppo di circolarità* di un socio "*s"* per un periodo di tempo "*dt*" l'insieme *GC(s,dt)* di negozi che *s* frequenta in modo significativo nel periodo *dt.* Si divida il dataset a disposizione in periodi di tempo dt_1, ..., dt_n consecutivi e di uguale ampiezza, scegliendoli in modo opportuno e motivato, e si estraggano i corrisponenti gruppi di circolarità per ogni socio. Nota: questo richiede di definire quando un socio "frequenta in modo significativo" un negozio, scelta importante che viene lasciata agli studenti;
4. il risultato del passo precedente è una sequenza di insiemi $GC(s,dt_1),..,GC(s,dt_n)$ per ogni socio *s*. Da questi estrarre dei *pattern di circolarità* frequenti, ovvero delle sequenze di gruppi o sotto-gruppi di circolarità che occorrono frequentemente tra i soci. Nota: una soluzione naturale consiste nell'usare pattern sequenziali, ottenendo pattern del tipo {neg.1, neg.2} --> {neg.3}. In certi casi, però, è interessante trovare pattern del tipo {neg.1, neg.2} in dt_1 --> {neg.3} in dt_2, ovvero associando strettamente ogni gruppo di circolarità al periodo di tempo in cui compare. Come si potrebbe realizzare con le tecniche di pattern mining di base, ovvero pattern sequenziali o itemset frequenti? Una volta discusso tale quesito, si estraggano i pattern di circolarità secondo la definizione che si preferisce (pattern sequenziali o l'approccio alternativo discusso), scegliendo e descrivendo poi i pattern ritenuti più significativi.

Progetto 4 (Inquinamento Pisa)

Sono a nostra disposizione i dati relativi alle seguenti fonti di informazione per la città di Pisa: (a) la mortalità dal 1990 ad oggi; (b) i ricoveri ospedalieri dal 1998 ad oggi; (c) anagrafe comunale, linkato ai precedenti; (d) analisi ISTAT relative alle sezioni di censimento della città; (e) livelli di inquinamento giornaliero della città; (f) dati meteo. Si indicano i seguenti obiettivi generali di analisi, che verranno meglio specificati in un succesivo incontro con l'esperto del dominio:

1. eseguire analisi per sezioni di censimento e/o circoscrizioni, su aree di malattie selezionate in bambini/giovani/anziani e uomini/donne. Nell'affrontare piccole aree geografiche, occorre tenere conto di alcune "anomalie" che possono alterare

le statistiche, quali la presenza in loco di case di riposo, domicili solo formali (es.: immigrati domiciliati presso il comune), ecc. Inoltre, si può tener conto della durata della residenza, laddove si cercano relazioni tra residenza e potenziale esposizione a sorgenti inquinanti di qualche genere;
2. valutare come aggregare le sezioni di censimento in unità omogenee per poi analizzare in esse i fenomeni sanitari;
3. simmetricamente al punto 2, analizzare i fenomeni sanitari separatamente per sezione di censimento, quindi raggruppare quelle che mostrano caratteristiche simili, estraendo un profilo per ogni cluster ottenuto.

Esperto del dominio: dott.ssa Maria Angela Vigotti (IFC-CNR Pisa)

Progetto 5A e 5B (Inquinamento Taranto A e B)

Sono a nostra disposizione i dati relativi alle seguenti fonti di informazione per la città di Taranto: (a) la mortalità dal 1998 al 2004 con link ai dati dell'anagrafe comunale (in particolare, le residenze sono scritte in modo corretto e uniformato); (b) la mortalità dal 2002 al 2004 con link all'anagrafe sanitaria (questa contiene problemi di uniformazione degli indirizzi); (c) i ricoveri ospedalieri dal 1998 al 2004, con link all'anagrafe sanitaria; (d) livelli di inquinamento giornaliero della città; (e) analisi ISTAT relative alle sezioni di censimento della città. Si indicano i seguenti obiettivi generali di analisi, che verranno meglio specificati in un succesivo incontro con l'esperto del dominio:

1. normalizzare lo stradario degli indirizzi riportati sull'anagrafe sanitaria, prendendo come modello di scrittura gli indirizzi dei deceduti del comune. Sull'anagrafe sanitaria 2002 e 2003, in particolare, ogni indirizzo si riferisce a 9 distretti (che successivamente sono diventati solo 4), quindi occorre riportare tutti gli indirizzi ai 9 distretti o eventualmente raggruppare gli indirizzi in nuove aree;
2. recuperare, se possibile, una mappatura delle sezioni di censimento di Taranto, così da mettere in relazione le analisi ISTAT con i dati di ricoveri e mortalità, e quindi fare una caratterizzazione socio-demografica delle aree di Taranto;
3. studiare le associazioni tra inquinamento giornaliero e ricoveri giornalieri in relazione alla vicinanza/lontananza da fonti inquinanti note (acciaieria, raffineria, cementificio) o arsenale militare (amianto, ecc.);
4. analisi, con strumenti di data mining, sui ricoveri per patologie selezionate e per gruppi di individui: genere (m,f), età (anziani, bambini o giovani), per identificare le aree o rioni a rischio per caratteristiche loro o per la loro vicinanza da fonti inquinanti.

Esperto del dominio: dott.ssa Maria Angela Vigotti (IFC-CNR Pisa)

Progetto 6A e 6B (Cirrosi epatica)

I dati in nostro possesso descrivono in modo parziale, sia a livello genetico che non, due popolazioni di individui: sani (circa 5000) e malati di cirrosi epatica (circa 500). Le variabili a nostra disposizione nel primo caso sono le seguenti: sesso, provincia di nascita, età, gruppo sanguigno, e coppie di espressioni di geni HLA (porzione del DNA comprendente il sistema immunitario): A,B,Drb1, ecc. Le coppie più complete sono A,B e Drb1, i due valori vengono uno dalla madre, uno dal padre, ed i valori mancanti indicano un valore uguale nei due genitori.

Il dataset relativo ai malati di cirrosi contiene informazioni simili: sesso (1=maschio, 2=femmina), per alcuni la provincia di nascita, età, coppie di espressioni dei geni A, B, Drb1 (qui i valori sopra il 90 indicano gene omozigote, ossia valore uguale in entrambi i genitori), ed un codice malattia (con 3 valori possibili: 1001, 1002, 1004 indicanti cirrosi da HCV, HBV o autoimmune)

Si indicano i seguenti obiettivi di analisi:

1. sul dataset degli individui sani, si cerchino pattern frequenti / regole associative sui geni, anche in relazione alle variabili sesso, età e provincia. Verificare quali di queste ultime tre variabili risulta essere in maggior rapporto con le caratteristiche genetiche osservate, ripetendo l'estrazione di pattern/regole separatamente su uomini e donne, su diverse fasce di età e su diverse province, confrontando i pattern ottenuti sui diversi segmenti di popolazione. Si tenga conto del fatto che la maggior parte degli individui qui descritti provengono dall'area pisana;
2. effettuare una analisi analoga sul dataset dei malati di cirrosi, tenendo conto del fatto che tali malati provengono essenzialmente da tutta Italia;
3. incrociare le due sorgenti di dati, cercando di estrarre (a) un classificatore che in base alle variabili a disposizione dica se un individuo è sano o malato; (b) un classificatore che determini l'area geografica di appartenenza; (c) altre eventiali.

Esperto del dominio: Michele Berlingerio (KDD Lab. ISTI-CNR Pisa)

Progetto 7A e 7B (Workflow mining)

Una azienda al disegno di apparecchiature meccaniche (specialmente pompe e motori) con strumenti CAD, fa uso di un sistema di gestione delle risorse che registra l'accesso ai diversi file creati e manipolati dai propri utenti. Il dataset contiene i log di tale software, e ogni record contiene: (a) l'operazione effettuata su un file (LOGOPEC); (b) la descrizione della operazione (LOGDESC); (c) timestamp dell'operazione (LOGDATE); (d) l'utente che effettua l'operazione (LOGUSER); (e) il gruppo dell'utente (LOGGROUP); (f) il ruolo dell'utente (LOGROLE); (g) file sul quale è effettuata l'operazione (LOGENAM).

Obiettivi del progetto:

1. pattern sequenziali sui file: trovare sequenze frequenti di operazioni effettuate su

uno stesso file, ripetendo l'esperimento dopo aver creato un generalizzazione delle operazioni che le riduca a 10-20 classi;

2. (Solo per progetto 7A) Usando le classi di operazioni individuate al passo precedente, calcolare la frequenza con cui ogni operazione (ovvero classe di operazioni) A effettuata su un file viene immediatamente seguita da una operazione B sullo stesso file, per tutte le possibili coppie (A,B). Nota: tra i valori di B si includa anche il valore "nessuna operazione", corrispondente ai casi in cui A è l'ultima operazione effettuata sul file. Rappresentare l'informazione così estratta in qualche forma grafica, possibilmente un grafo delle transizioni A-->B[1];
3. estrarre clustering di utenti che mostrano comportamenti simili. In particolare, questo richiede di modellare nel modo più preciso possibile tali comportamenti in forma di semplici variabili, a partire da sequenze di operazioni. Esempi di tali variabili (alcune semplici, altre molto complesse) sono: numero di operazioni effettuate, quali sono le operazioni effettuate, quanto dura una sessione di lavoro, quali tempi intercorrono tra una operazione e l'altra, quali pattern frequenti occorrono nella storia dell'utente, ecc. Quindi, fornire un profilo per ogni cluster;
4. predizione del ruolo degli utenti: utilizzare le variabili costruite nei punti precedenti per costruire un classificatore in grado di distinguere gli utenti con ruolo uguale a view dagli altri.

Esperto del dominio: Fabio Pinelli (KDD Lab. ISTI-CNR Pisa)

1 Una possibilità (non l'unica) consiste nell'usare software come "dot" del package "graphviz".