Università degli Studi di Milano
Corso di laurea Magistrale in Informatica

Progetto di Gestione dell'Informazione

Data mining su dati finanziari

DOCENTE
Andrea G.B. Tettamanzi

PROGETTO DI
Guido Lena Cota - 773958

Anno accademico 2010/2011

Indice

Capitolo 1

Introduzione

Una delle più interessanti conseguenze dell'evoluzione tecnologica degli ultimi decenni la disponibilità di un'impressionante numero di dati in modo semplice ed economico. Tale fenomeno è stato incoraggiato da almeno quattro fattori:

- Progressi nei sistemi hardware di elaborazione e memorizzazione, sia in termini di costo che di efficienza;
- Sviluppo di sistemi software ottimizzati per la gestione di grandi volumi di informazioni, quali database e altre strutture dati;
- Perfezionamento di metodi automatici di rilevazione dei dati;
- Informatizzazione di molte banche dati, che fino a poco tempo fa erano consultabili solo in formato cartaceo.

L'accesso a una simile mole di informazioni sta ribaltando la prospettiva del metodo scientifico tradizionale: i dati non sono più usati unicamente per dimostrare ipotesi, ma diventano essi stessi oggetto di studio. Ciò è reso possibile dagli strumenti informatici e dalle tecniche che presenterò nei prossimi paragrafi, senza le quali sarebbe impensabile o quanto meno poco conveniente un simile approccio. Si tratta infatti di estrarre informazioni da numerosi dati non strutturati, spesso provenienti da settori diversi (IT, finanza, ricerca, ...) e quindi poco integrati.

1.1 Estrazione della conoscenza

Con il termine *Knowledge Discovery in Database* (KDD) si intende il processo di estrazione di conoscenza da grandi masse di dati grezzi, attraverso la scoperta automatica di regolarità o relazioni (*pattern*) non note a priori [1]. Si tratta di una disciplina recente, che si pone come punto di congiunzione tra

diverse aree scientifiche, dalla statistica all'intelligenza artificiale alla gestione di database.

Il *Data Mining* è una delle fasi del KDD, anche se viene spesso confuso con l'intero processo di estrazione della conoscenza. Per comprendere meglio il suo ruolo, riporterò i cinque passi della *Knowledge Discovery in Database* individuati in [3] e schematizzati in figura 1.1:

1. Selezione, che estrae dai dati grezzi quelli considerati rilevanti per le analisi. I criteri di selezione dipendono dal problema che si intende affrontare, il che rende necessaria una buona conoscenza dello scenario di lavoro;

2. Pre-elaborazione, che tenta di ridurre i dati errati, incompleti o rumorosi (*data cleaning*), e di integrare dati provenienti da più fonti risolvendo le inconsistenze (*data integration*);

3. Trasformazione, che prepara o consolida i dati in forme più appropriate per gli algoritmi di analisi. Esempi comuni di trasformazione sono la normalizzazione, la discretizzazione e la riduzione;

4. *Data Mining*, che attraverso l'applicazione di algoritmi specifici permette di estrarre modelli significativi dai dati, fornendo così informazioni efficaci rispetto agli obiettivi della ricerca. E' caratterizzato da due fasi, una di esplorazione dei dati e una di generazione (automatica o interattiva) dei *pattern*;

5. Interpretazione e valutazione dei modelli identificati. Qualora questi ultimi non fossero soddisfacenti, è possibile tornare alle fasi precedenti e rivedere alcuni passaggi.

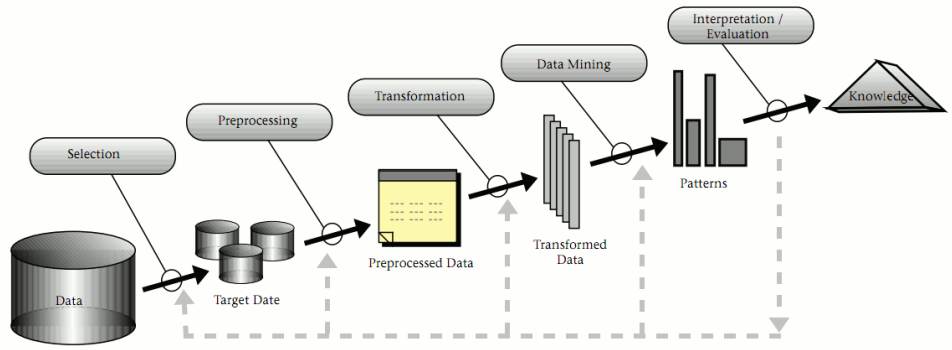

Figura 1.1: Fasi del processo di estrazione della conoscenza. Fonte: [2]

1.2 Tecniche di Data Mining

Il *Data Mining* rende possibile l'estrazione di informazioni utili al supporto decisionale, lavorando su dataset di grandi dimensioni. Esistono diverse tecniche, ognuna caratterizzata da un algoritmo e da una struttura di conoscenze associata.

Le tecniche di *Data Mining* possono essere classificate in due categorie: ad apprendimento supervisionato o non supervisionato. Gli algoritmi del primo tipo operano su esempi, ovvero dati di cui è già nota la corretta classificazione; le tecniche non supervisionate non possono invece contare su alcuna conoscenza dei loro contenuti.

L'implementazione di strategie *Data Mining* supervisionate permette la risoluzione di classificazione, stima e previsione, mentre le non supervisionate consentono clusterizzazioni e associazioni.

1.2.1 Regole di associazione

Per associazione si intende una relazione o una dipendenza significativa tra attributi che presentano caratteristiche comuni. Grazie al *Data Mining* è possibile quindi scoprire legami non prevedibili tra elementi diversi, o relazioni causa-effetto non ancora tenute in considerazione.

Tali regole vengono formalmente espresse come implicazioni del tipo $X \Rightarrow Y$, dove X e Y sono insiemi di oggetti disgiunti, chiamati *itemset*. In particolare X è detto l'antecedente della regola ed Y il conseguente. La regola è verificata se esiste almeno un record nel dataset che contenga gli oggetti dell'insieme $X \cup Y$.

Il *supporto* di una regola di associazione rappresenta la frequenza della stessa all'interno del dataset, mentre la *confidenza* misura la validità dell'implicazione logica. Più i loro valori sono alti e più la regola è forte.

L'obiettivo di un processo di estrazione di regole associative è generare tutte le regole che soddisfano un livello minimo di supporto e confidenza, specificato dall'utente.

1.2.2 Algoritmo Apriori

L'algoritmo Apriori è stato proposto in [5] da Agrawal e Srikant, ed è considerato il metodo classico di *Data Mining* per l'estrazione di regole di associazione.

L'algoritmo ha due fasi:

1. Ricerca di tutti gli *itemset* frequenti, cioè quelli che si ripropongono più spesso nel dataset. La soglia minima di frequenza è definita dall'utente, e non è altro che il supporto minimo ricercato;

2. Generazione delle regole di associazione a partire dai candidati individuati nel passo precedente.

Per ridurre lo spazio di ricerca, la prima fase viene affrontata partendo dall'assunzione che ogni sottoinsieme di un *itemset* frequente è esso stesso un *itemset* frequente. Vale ovviamente anche il viceversa: se un *itemset* non è frequente allora nessuno dei suoi sovrainsiemi lo sarà, rendendo quindi superfluo considerare gli insiemi che lo includono. La generazione dei candidati seguirà dunque un approccio di tipo "bottom-up", in cui si inizia da insiemi frequenti composti da un unico elemento, e si costruisce iterativamente *itemset* più grandi partendo da quelli frequenti trovati all'iterazione precedente.

La seconda fase dell'algoritmo consiste nella generazione delle regole di associazione. Per ogni *itemset* frequente (I), si forma una regola $X \Rightarrow Y$ in cui il conseguente (Y) è un sottoinsieme di I, mentre l'antecedente (X) è formato dai valori di I che non compaiono in Y. Ad esempio se $I = \{A, B, C\}$, una regola potrebbe essere $\{A, B\} \Rightarrow C$. La confidenza di tale regola può essere calcolata come il rapporto tra i supporti degli *itemset* X e Y, e se supera il valore minimo definito dall'utente, allora sarà tenuta; tutte le altre vengono scartate.

1.3 Organizzazione del progetto

Scopo del progetto è ricavare regole associative a partire da una serie di informazioni finanziarie.

Nel Capitolo 2 sarà presentato il caso di studio, mentre nel Capitolo 3 e 4 saranno descritte le fasi operative del progetto e commentati i risultati.

Capitolo 2

Caso di studio

Un mercato è il luogo in cui avvengono gli scambi economici del sistema economico di riferimento. Il mercato finanziario è dato dall'insieme delle operazioni aventi per oggetto strumenti finanziari a medio e lungo termine (da 18 mesi a 5 anni), come azioni, Titoli di Stato o contratti *futures*. Al suo interno si possono distinguere altri mercati, tra cui quello monetario, obbligazionario, azionario e valutario. Il mercato monetario è il luogo della domanda e dell'offerta di moneta e di titoli di credito a breve scadenza (depositi postali, BOT, fondi comuni di investimento, ...), ed ha come obiettivo primario la gestione della liquidità.

La finanza è un settore pesantemente informatizzato, caratterizzato dalla produzione continua di enormi quantità di dati. Non sorprende dunque che già dagli anni '80 la comunità finanziaria si mostrasse interessata al supercalcolo e al *Data Mining*, una disciplina che si stava appena affacciando in ambito marketing [4]. Il problema di valorizzazione dei derivati[1] e di misurazione dei rischi coinvolgono infatti un universo di titoli e/o contratti molto vasto, delineando uno scenario di applicazione ideale per le strategie di estrazione della conoscenza.

In questo progetto applicherò tecniche *Data Mining* per individuare *pattern* e similarità nelle serie storiche di un insieme di indicatori finanziari. Le applicazioni sono estremamente interessanti, perché l'eventuale scoperta di relazioni di causalità robuste, o meglio ancora di similarità strutturali, permetterebbe di prevedere gli andamenti futuri degli indici in modo molto più accurato.

Nei prossimi paragrafi saranno presentati gli indicatori che compongono il dataset. Uno dei principali criteri di scelta è che abbiano un numero sufficientemente ampio di dati, motivo per cui sono stati considerati solo indicatori con una pubblicazione quotidiana dei valori. Altre informazioni interessanti come tassi di inflazione o debiti pubblici nazionali sono stati

[1] Ogni contratto o titolo finanziario il cui prezzo è basato sul valore di mercato di altri beni (azioni, indici, valute, tassi ecc.).

dunque ignorati non per importanza, ma per scarsità di dati (generalmente agglomerati in mensilità o - più frequentemente - annualità).

2.1 Commodity

Le *commodity* sono beni che vengono offerti sul mercato senza differenze qualitative, grazie a un elevato livello di standardizzazione e controllo. Queste caratteristiche rendono agevoli le negoziazioni sui mercati finanziari, e costituiscono un'attività sottostante per vari tipi di strumenti derivati, in particolare per i *future*[2].

Generalmente le *commodity* sono prodotti agricoli, energetici, metalli o prodotti di base non lavorati (ad esempio caffè o zucchero).

2.1.1 Costo del petrolio

Negli scambi commerciali il petrolio viene tradizionalmente quotato rispetto a due *benchmark*: il WTI e il Brent, le cui piazze principali sono rispettivamente il NYMEX di New York e l'IntercontinentalExchange (ICE) di Atlanta. In entrambi i casi il prezzo è espresso in dollari americani per barile (circa 159 litri), con un quantitativo minimo di 1000 barili a transazione.

Il WTI (*West Texas Intermediate*) fa da riferimento alle transazioni interne agli Stati Uniti, mentre il Brent al 60% degli scambi internazionali di greggio. Ovviamente esistono altri benchmark (Dubai, DME Oman, ...), ma sono meno utilizzati.

I dati che andranno a formare il dataset sono presi dallo *Statistical Data Warehouse* della Banca Centrale Europea, raggiungibile a questo indirizzo web: http://sdw.ecb.europa.eu/

2.1.2 Costo dei metalli

I metalli preziosi sono considerati ottimi beni rifugio per diversi fattori: rarità, inalterabilità nel tempo, scarsa correlazione con gli andamenti azionari e obbligazionari, facilità di trasporto, disponibilità di un valore di riferimento ufficiale. Per tutti questi motivi sono tra le *commodity* più scambiate nei mercati finanziari.

I metalli di cui monitorerò i prezzi sono oro, argento e platino. Le quotazioni sono espresse in dollari statunitensi per oncia, e si considereranno quelle fissate due volte al giorno dalla Borsa di Londra (*London Fixing*).

La fonte dati è il portale http://www.kitco.com/

[2]Contratti a termine standardizzati per poter essere negoziati facilmente in Borsa.

2.2 Tassi di cambio

Il tasso di cambio è il valore di un'unità di valuta in termini di un'altra valuta, quindi il suo prezzo espresso in un'altra moneta. Come ogni altro bene, il suo valore subisce variazioni in funzione di cambiamenti del mercato e delle leggi della domanda e dell'offerta.

I tassi di cambio considerati sono Euro-Dollaro Statunitense (EUR-USD) e Euro-Sterlina Inglese (EUR-GBP). I valori sono presi dalla banca dati della Banca Centrale Europea.

2.3 Tassi di interesse

I principali tassi di interesse del mercato monetario Europeo sono Euribor (*EURo Inter Bank Offered Rate*) ed Eonia (*EURo OverNight Interest Average*). In entrambi i casi si utilizzeranno i valori pubblicati dalla BCE.

2.3.1 Euribor

L'Euribor indica il tasso di interesse medio delle transazioni finanziarie in Euro tra le principali banche europee (oltre 50). Il suo valore dipende dalla durata del prestito (da 1 a 12 mesi), indipendentemente dall'ammontare del capitale. In questo caso di studio ho considerato il tasso a scadenza mensile, perché il più influenzato dai cambiamenti quotidiani del mercato.

L'Euribor è un indicatore del costo del denaro a breve termine, ed è spesso usato come tasso di riferimento per il calcolo di interessi variabili, come quello dei mutui. Il suo valore è pubblicato ogni giorno alle 11.00 dalla *European Banking Federation* (EBF).

2.3.2 Eonia

L'Eonia è il tasso di interesse medio al quale una selezione di banche europee si concede reciprocamente prestiti in euro per un periodo di 1 giorno. È il primo tasso a risentire delle decisioni di politica monetaria.

2.4 Indici azionari

Gli indici azionari forniscono una sintesi del valore di un portafoglio di titoli azionari, e sono un importante indicatore dell'andamento globale di borsa e finanza.

La sorgente dati degli indici che seguiranno è il sito http://wikiposit.org/

2.4.1 FTSE-100

FTSE sta per *Financial Times Stock Exchange*, e l'FTSE-100 è l'indice azionario delle 100 società più capitalizzate quotate al *London Stock Exchange* (circa l'80% della capitalizzazione del suo mercato). Le società vengono aggiornate trimestralmente dall'FTSE Group in piena autonomia, e attualmente ne fanno parte aziende come Lloyds TSB, Rolls-Royce plc, Unilever e Vodafone.

L'indice è quotato a partire dal 3 gennaio 1984, motivo per cui è stato preferito all'italiano FTSE MIB, la cui serie storica comincia il primo giugno 2009.

2.4.2 NASDAQ-100

Analogamente all'FTSE-100, il NASDAQ-100 è l'indice di borsa delle maggiori 100 imprese non finanziarie quotate al mercato NASDAQ, composto dai principali titoli tecnologici della borsa statunitense (IBM, Microsoft, Apple, ...) e straniera (Ericsson, Logitech, Ryanair, ...).

Il NASDAQ-100 (NDX) è un indice ponderato, in cui il peso delle società che lo compongono è basato sulla loro capitalizzazione di mercato, con alcune regole per tener conto delle influenze delle aziende maggiori.

2.4.3 S&P GSCI

L'S&P GSCI è un indice introdotto nel 2007 dalla banca d'affari Goldman Sachs, ed oggi di proprietà dell'agenzia di rating Standard & Poor's. Viene calcolato come la media pesata di 24 diversi contratti *future*, dal petrolio ai gas naturali, dai metalli preziosi ai prodotti agricoli; per questo motivo è usato come riferimento per gli investimenti nel mercato delle *commodity*.

2.4.4 S&P 500

L'indice S&P 500 è stato introdotto da Standard & Poor's nel 1957 per seguire l'andamento azionario delle 500 aziende statunitensi a maggiore capitalizzazione. Il peso attribuito a ciascuna di esse è direttamente proporzionale al suo valore di mercato.

È l'indice più usato per misurare l'andamento del mercato azionario USA, ed è riconosciuto come *benchmark* per le performance di portafoglio.

L'S&P 500 ha diversi simboli, tra cui GSPC - usato in questo progetto - e SPX.

Capitolo 3

Fasi operative

Nel capitolo saranno presentate tutte le operazioni necessarie per l'estrazione di regole associative dal dataset.

3.1 Trattamento dei dati

I dati selezionati per l'indagine sono gli indicatori elencati nel capitolo 2.

Le relative serie storiche, considerate tra il 4 gennaio 1999 e il 15 novembre 2011, sono salvate in diversi fogli di lavoro della cartella LibreOffice Calc `datasetGI.ods` allegata alla relazione. Ogni serie è composta da due colonne: la data di riferimento e il valore dell'indicatore.

3.1.1 Pre-elaborazione

La fase di pre-elaborazione dei dati è un passaggio obbligato per il trattamento di grandi moli di informazioni, in particolare quando bisogna accorpare dati provenienti da fonti diverse.

Un primo passo di *data cleaning* è stato operato per correggere le inconsistenze dei dati, dovute principalmente a un'errata interpretazione del simbolo "." da parte di Calc. La maggior parte delle serie storiche sono state infatti prese da database stranieri, in cui la notazione per i numeri decimali prevede l'uso del punto e non della virgola. Per alcuni record, non per tutti, ciò è stato interpretato come indicatore delle migliaia, da cui le inconsistenze. Visualizzando graficamente i dati l'errore è ancora più evidente, data la totale assenza di valori intermedi tra due picchi che differiscono di ben tre ordini di grandezza. Per risolvere l'inconsistenza è bastato creare una formula come quella mostrata in figura 3.1: se il valore dell'indicatore supera una certa soglia (in questo caso 1000, ma in generale potrebbe essere la media dei valori), allora viene diviso per 1000, altrimenti viene lasciato com'è.

Il secondo passo di pre-elaborazione è stato l'integrazione dei dati appartenenti a serie storiche diverse in un unico dataset coerente. Un semplice

Figura 3.1: Correzione delle inconsistenze nelle serie di dati.

merge delle colonne dei valori in funzione delle date non è sufficiente, perché le date di riferimento non sono nello stesso formato e non sono sempre uguali.

Il problema del formato è stato risolto utilizzando la formula di Calc *DATEVALUE*, che restituisce il numero seriale della data rappresentata in una cella come testo. Ripetendo l'operazione per ogni serie storica si è ottenuta una notazione univoca per tutte. Il secondo problema ha richiesto l'impiego della *VLOOKUP*, che verifica la presenza di un certo valore in un insieme di colonne, e se lo trova restituisce il contenuto di una delle celle adiacenti. In figura 3.2 è mostrato un esempio di applicazione della formula.

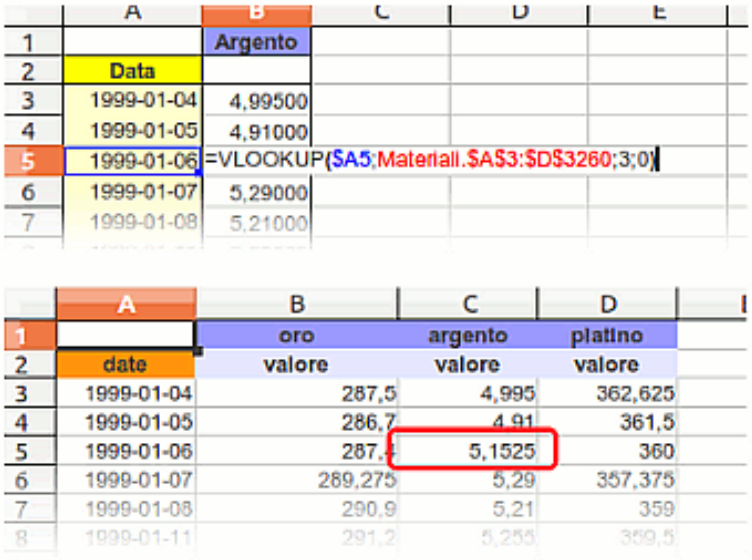

Figura 3.2: Integrazione delle serie storiche provenienti da diversi fogli di lavoro. La formula dell'immagine in alto può essere tradotta come: cerca nelle prime 4 colonne del foglio di lavoro `Materiali` la data *1999-01-06*, e se la trovi riporta il valore della terza colonna.

Il risultato dell'integrazione è riportato nel foglio di lavoro `Dataset` del file Calc allegato alla relazione.

L'ultimo passaggio di pre-elaborazione è stato la risoluzione delle nuove inconsistenze e dei dati mancanti, inevitabile conseguenza dell'integrazio-

ne tra serie storiche diverse. La politica risolutiva è stata usare la media tra i primi dati noti che comprendono quello mancante, così da preservare il significato e il trend locale delle serie. Ciò si traduce operativamente nell'applicazione della formula mostrata in figura 3.3.

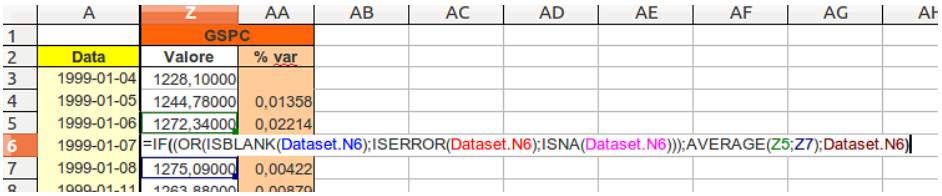

Figura 3.3: Correzione dei dati mancanti originati dall'integrazione delle serie dati.

Le colonne dei valori del foglio di lavoro `Dataset-clean` sono il risultato di tutti i passaggi di pre-elaborazione, e saranno il punto di partenza per la fase successiva di trasformazione.

3.1.2 Trasformazione

I valori degli indicatori delle serie storiche sono molto differenti tra loro, nonostante l'integrazione e la correzione delle inconsistenze. A tale proposito è stata effettuata una sorta di normalizzazione, studiando le variazioni percentuali piuttosto che il valore effettivo. Ho dunque creato una nuova colonna di dati per ogni indicatore, nelle cui celle viene calcolata la variazione di valore (positiva o negativa) rispetto all'istanza precedente.

Secondo e ultimo passo è stato la discretizzazione dei valori in categorie, perché gli algoritmi di ricerca di regole di associazione operano solo su dati discreti. Per completezza di analisi, ho considerato due casi: il Caso A ha 5 categorie ed è salvato nel file `disc5.csv`, mentre il Caso B ne ha 3 ed è salvato in `disc3.csv`. La formula di discretizzazione utilizza soglie ricavate dai percentili della distribuzione dati.

Le categorie del Caso A sono così composte:

1. $=$, stabilità sostanziale: tra il 45-simo e 55-simo percentile. Per ogni indicatore ho verificato - empiricamente - che il passaggio da valori negativi a positivi avvenga in questo intervallo;

2. $<<$, variazione molto negativa: tutti i valori fino al 22,5-simo percentile. Il valore è stato determinato come punto medio tra lo 0 e la soglia del 45 individuata nella categoria precedente;

3. $<$, variazione negativa: tra il 22,5-simo e 45-simo percentile;

4. $>$, variazione positiva: tra il 55-simo e 77,5-simo percentile;

5. $>>$, variazione molto positiva: i valori oltre il 77,5-simo percentile.

Nel Caso B le categorie sono state discretizzate secondo i seguenti criteri:

1. =, stabilità sostanziale: tra il 40-simo e 60-simo percentile;
2. <, variazione negativa: tutti i valori fino al 40-simo percentile;
3. >, variazione positiva: tutti i valori oltre il 60-simo percentile.

3.2 Data-mining

Per la ricerca di regole di associazione nel dataset ho utilizzato il software open-source *Weka* (http://www.cs.waikato.ac.nz/ml/weka/), sviluppato nell'università di Waikato in Nuova Zelanda e parte integrante del framework Pentaho per la business intelligence. Il programma offre una vasta collezione di algoritmi di *Data Mining*, tra cui l'Apriori presentato nel capitolo introduttivo del progetto.

All'avvio Weka fa scegliere tra quattro diverse modalità di lavoro: Explorer, Knowledge Flow, Experimenter e Simple CLI. Per il progetto ho utilizzato la modalità Explorer, che permette di elaborare e visualizzare i dati in modo esaustivo attraverso un'interfaccia grafica intuitiva.

Per caricare i dati bisogna cliccare sul pulsante *Open File* della sezione `Preprocess`. Nella finestra appariranno tutti gli attributi (le serie storiche), una serie di informazioni statistiche, e dei grafici a barre che riportano l'andamento dei valori (vedi figura 3.4).

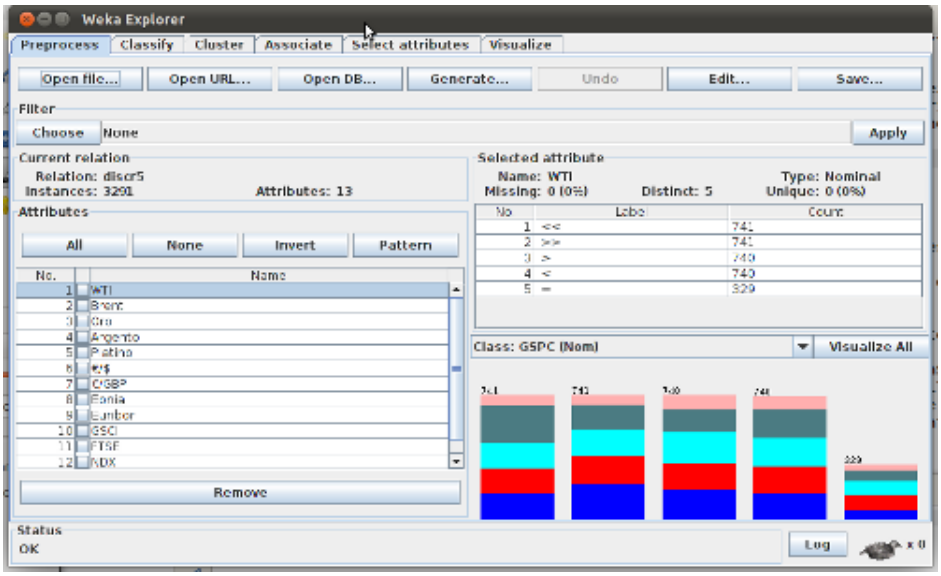

Figura 3.4: Sezione *Preprocess* di Weka.

Per applicare l'algoritmo Apriori bisogna portarsi nella sezione `Associate` e selezionarlo tra le diverse modalità di estrazione regole. Facendo clic sul testo accanto il pulsante `Choose` apparirà la finestra di configurazione dei parametri dell'algoritmo (vedi figura 3.5). I parametri settati sono:

- *lowerBoundMinSupport*: supporto minimo che devono soddisfare le regole per non essere scartate;
- *delta*: unità di decremento del supporto corrente ad ogni iterazione;
- *minMetric*: limite minimo di confidenza delle regole trovate;
- *numRules*: numero massimo di regole da generare.

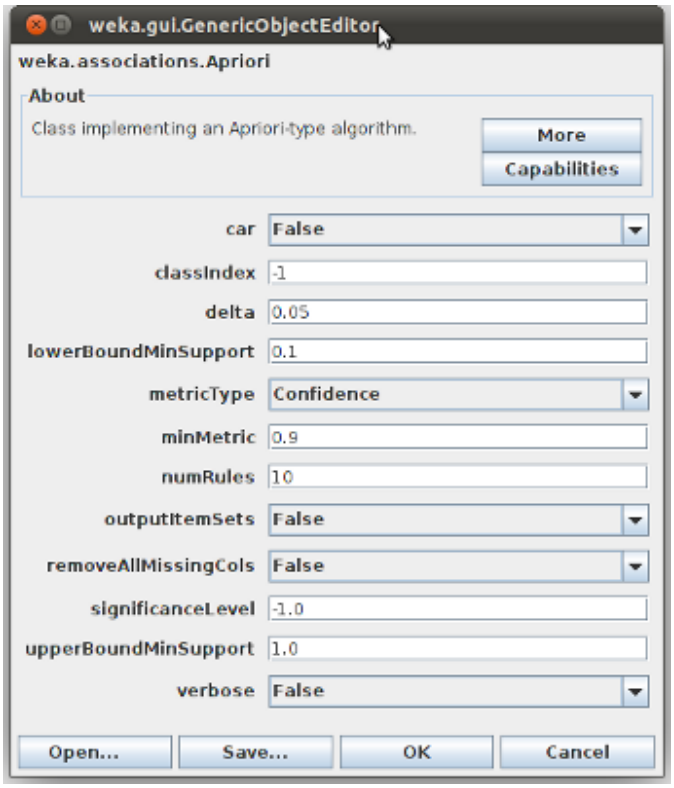

Figura 3.5: Configurazione dei parametri dell'algoritmo Apriori in Weka.

Cliccando sul pulsante `Start` l'algoritmo inizia a cercare le prime *numRules* regole che superano la soglia *minMetric* di confidenza. La ricerca parte considerando le regole che hanno un supporto del 100%, che ad ogni ciclo iterativo viene decrementato di *delta* fino a raggiungere il *lowerBoundMinSupport*.

I risultati sono pubblicati nel riquadro *Associator output*, e saranno commentati nel prossimo capitolo.

Capitolo 4

Risultati

4.1 Caso A

Per il Caso A il dataset è stato discretizzato in cinque categorie. I parametri di configurazione dell'algoritmo Apriori in Weka sono:
lowerBoundMinSupport = 0,1; *minMetric* = 0,5.

Nella seguente tabella sono riportate le prime 20 regole di associazione trovate dall'algoritmo, ordinate per livello di confidenza. I valori della colonna ID sono usati per fare riferimento alle regole nei commenti.

ID	Regole di associazione	Supporto	Confidenza
1	WTI << Brent << 385 ==> GSCI << 347	10,54	90
2	WTI >> Brent >> 379 ==> GSCI >> 333	10,12	88
3	WTI << 741 ==> GSCI << 585	17,78	79
4	WTI >> 741 ==> GSCI >> 574	17,44	77
5	GSPC << 741 ==> NDX << 564	17,14	76
6	NDX << 741 ==> GSPC << 564	17,14	76
7	GSPC >> 741 ==> NDX >> 553	16,80	75
8	NDX >> 741 ==> GSPC >> 553	16,80	75
9	Eonia = 914 ==> Euribor = 633	19,23	69
10	Oro > 740 ==> Euribor = 467	14,19	63
11	Argento << 741 ==> Oro << 464	14,10	63
12	Oro << 741 ==> Argento << 464	14,10	63
13	FTSE > 740 ==> Euribor = 461	14,01	62
14	Brent < 740 ==> Euribor = 459	13,95	62
15	FTSE < 740 ==> Euribor = 459	13,95	62
16	GSPC < 740 ==> Euribor = 456	13,86	62
17	GSCI > 740 ==> Euribor = 455	13,83	61
18	WTI > 740 ==> Euribor = 453	13,76	61
19	€/$ > 740 ==> Euribor = 453	13,76	61
20	NDX > 740 ==> Euribor = 452	13,73	61

Le prime quattro regole in tabella mostrano una correlazione ad alta confidenza tra i prezzi del greggio e l'indicatore GSCI del mercato delle *commodity*. In particolare, le regole 1 e 2 dicono che a fronte di una forte variazione del WTI e del Brent, il GSCI manterrà lo stesso andamento. Le regole 3 e 4 confermano il comportamento anche quando l'antecedente è composto da uno solo dei due prezzi del petrolio.

Le regole dalla 5 alla 8 mostrano una dipendenza analoga tra l'indice azionario S&P 500 e il Nasdaq-100. Si tratta di un altro risultato prevedibile, dal momento che entrambi monitorano gli andamenti delle società statunitensi più capitalizzate.

Un'altra correlazione è individuata dalle regole 10 e 11, che mostrano comportamenti simmetrici nelle variazioni dei prezzi di oro e argento.

Infine, dalle rimanenti regole riportate in tabella si evince l'indipendenza del tasso di interesse Euribor a scadenza mensile rispetto alle variazioni di diversi indicatori. L'Euribor rimane infatti sempre stabile a fronte di diminuzioni o aumenti più o meno sostenuti degli indici azionari o dei prezzi delle *commodity*.

4.2 Caso B

Nel Caso B il dataset è stato discretizzato in tre categorie. I parametri di configurazione dell'algoritmo Apriori sono un po' più esigenti, perché con la diminuzione del numero di classi discrete mi aspetto maggiori probabilità di *itemset* frequenti:

$lowerBoundMinSupport = 0{,}2$; $minMetric = 0{,}5$.

Nella tabella riportata nella pagina successiva sono elencate le prime 20 regole di associazione trovate dall'algoritmo, ordinate per confidenza.

Tutte le regole che contengono informazioni sul greggio non fanno altro che confermare la dipendenza già emersa analizzando il Caso A, ovvero che gli andamenti del GSCI seguono coerentemente quelli di WTI e Brent. Stesso discorso anche per la correlazione tra Nasdaq-100 ed S&P 500. I livelli di confidenza e supporto di queste regole sono maggiori rispetto al caso precedente, rendendole di fatto ancora più attendibili.

In questo caso si rafforza poi il legame strutturale tra gli indicatori dei metalli preziosi, che aumentano e diminuiscono in modo coordinato.

4.3 Conclusioni

L'applicazione dell'algoritmo Apriori sugli indicatori finanziari scelti non ha prodotto regole forti di particolare interesse. Esse infatti non fanno altro che confermare relazioni più o meno strutturali piuttosto intuitive, senza rivelare nuove dipendenze impreviste.

ID	Regole di associazione	Supporto	Confidenza
1	WTI > Brent > 818 ==> GSCI > 744	22,61	91
2	WTI < Brent < 825 ==> GSCI < 741	22,52	90
3	FTSE < NDX < 740 ==> GSPC < 660	20,05	89
4	Brent > GSCI > 855 ==> WTI > 744	22,61	87
5	Brent < GSCI < 857 ==> WTI < 741	22,52	86
6	GSCI > 1317 ==> WTI > 1102	33,49	84
7	WTI > 1317 ==> GSCI > 1102	33,49	84
8	FTSE < GSPC < 789 ==> NDX < 660	20,05	84
9	WTI < 1316 ==> GSCI < 1099	33,39	84
10	GSPC < 1316 ==> NDX < 1080	32,82	82
11	NDX < 1316 ==> GSPC < 1080	32,82	82
12	GSPC > 1317 ==> NDX > 1061	32,24	81
13	NDX > 1317 ==> GSPC > 1061	32,24	81
14	Oro < Platino < 845 ==> Argento < 677	20,57	80
15	Oro > Platino > 828 ==> Argento > 662	20,12	80
16	Oro > Argento > 925 ==> Platino > 662	20,12	72
17	Oro < Argento < 949 ==> Platino < 677	20,57	71
18	Argento > 1317 ==> Oro > 925	28,11	70
19	Oro > 1317 ==> Argento > 925	28,11	70
20	Platino < 1316 ==> Oro < 845	25,68	64

Tuttavia, abbassando leggermente le soglie minima di confidenza e supporto, emergono informazioni e comportamenti più interessanti. Ne sono un esempio le correlazioni tra andamenti dei tassi di cambio Euro-dollaro statunitense ed Euro-sterlina inglese, e contemporaneamente tra indici azionari americani (NASDAQ-100) ed Europei (FTSE-100). Entrambe le regole hanno una confidenza superiore al 60%, e sono quindi sufficientemente credibili per confermare una volta di più la stretta interdipendenza dei mercati nel moderno contesto globale

Bibliografia

[1] Dulli S., Furini S., Peron E., *Data mining: Metodi e strategie*, Springer Verlag.

[2] Fayyad U. et al. (1996), *From Data Mining to Knowledge Discovery in Databases*, AI Magazine (Vol 17, No 3), 37-54.

[3] Brachman R., Anand T. (1996), *The Process of Knowledge Discovery in Databases: A Human-Centered Approach*, In Advances in Knowledge Discovery and Data Mining, 37–58, eds. Fayyad, Piatetsky-Shapiro, Smyth e Uthurusamy. Menlo Park, Calif.: AAAI Press.

[4] Focardi S.M. (2001), *Clustering delle serie storiche economiche: Applicazioni e questioni computazionali*, The Intertek Group.

[5] Agrawal R., Srikant R. (1994), *Fast Algorithms for Mining Association Rules*, Proceedings of 20th International Conference on Very Large Data Bases