# KDB2000: UNO STRUMENTO DI SUPPORTO AL PROCESSO DI SCOPERTA DI CONOSCENZA NELLE BASI DI DATI

Donato Malerba, Annalisa Appice, Antonia Bellino, Michelangelo Ceci, Domenico Pallotta
Dipartimento di Informatica
Università di Bari
Via Orabona, 4 – 70126 Bari
{Malerba, Appice, Bellino, Ceci, Pallotta} @ di.uniba.it

## ABSTRACT.

*The automated discovery of knowledge in databases is becoming increasingly important as the worlds wealth of data continues to grow exponentially. Knowledge discovery systems face challenging problems from real world databases which tend to be dynamic, incomplete, redundant, noisy, sparse and very large. This paper describes KDB2000, a Data Mining tool which assists a user in carrying out the complex knowledge discovery process.*

**Key Words: KDD, Data Mining.**

*Con la crescita esplosiva nelle capacità di generare e collezionare dati cresce l'importanza di automatizzare il processo di scoperta di conoscenza nelle basi di dati. I sistemi per la scoperta di conoscenza devono affrontare i problemi tipici dei dati disponibili nelle basi di dati del mondo reale. Infatti, questi sono spesso incompleti, dinamici, ridondanti, incerti. Questo articolo offre una panoramica del processo KDD (knowledge discovery in databases). In particolare, si descrive KDB2000, un sistema di Data Mining che assiste l'utente in tutte le fasi del processo KDD.*

## 1. IL PROCESSO KDD

Il termine *Knowledge Discovery in Databases* (*KDD)*, coniato nel 1989, si riferisce all'intero processo, interattivo ed iterativo, di scoperta della conoscenza che consiste nell'identificazione di *relazioni tra dati* che siano *valide*, *nuove*, potenzialmente *utili* e *comprensibili* [1]. Per una migliore comprensione di tale definizione si esaminano in dettaglio i concetti in essa coinvolti.

- I *dati* sono una collezione di fatti *F* (per esempio tuple di una tabella di un database relazionale).
- Una *relazione* (o modello o *pattern*) è un'espressione *E* in un linguaggio *L* che descrive fatti in un sottoinsieme *FE* di *F*. Ovviamente una relazione deve essere più semplice rispetto ad un dato criterio di semplicità della enumerazione di tutti i fatti in *FE*.
- Un *processo* di scoperta della conoscenza è un insieme di attività che coinvolgono la preparazione dei dati, la ricerca di relazioni, la valutazione e il raffinamento della conoscenza estratta. Si assume che il processo sia non banale, cioè che le relazioni scoperte non siano già note.
- Le relazioni scoperte sono *valide* se valgono, con un certo grado di certezza, anche su dati diversi da quelli usati per la scoperta delle stesse. Individuare un grado di certezza è essenziale per stabilire quanta fiducia si può riporre nel sistema e nella scoperta effettuata. Il concetto di certezza coinvolge diversi fattori quali l'integrità dei dati, la dimensione del campione su cui è effettuata la scoperta o la quantità di conoscenza già disponibile.
- Le relazioni scoperte devono essere *nuove* almeno per il sistema. La novità può essere misurata rispetto ai cambiamenti nei dati (confrontando i valori correnti con quelli precedenti o quelli attesi) o nella conoscenza (cioè come una nuova scoperta è collegata a quella precedente).
- Le relazioni dovrebbero potenzialmente condurre a delle azioni *utili*. Per esempio la scoperta di una dipendenza fra articoli acquistati da uno stesso cliente in un supermercato potrebbe attivare opportune strategie di marketing.
- Le relazioni scoperte devono essere *comprensibili* agli utenti per facilitare una migliore comprensione dei dati coinvolti. Poiché è difficile misurare precisamente "la comprensibilità di un pattern" spesso si ricorre a misure surrogate di semplicità sintattica/semantica.

Il processo KDD può essere scomposto in più fasi tra le quali il Data Mining riveste un ruolo centrale[1]. Il primo passo del KDD, ossia *individuare gli obiettivi dell'utente finale*, ha molto in comune con le fasi iniziali di un qualunque progetto commerciale, durante le quali si individuano gli obiettivi di business da raggiungere. In questa fase è importante la collaborazione che si stabilisce tra gli analisti del business e gli analisti dei dati, i quali devono comprendere il dominio applicativo ed estrarre la conoscenza già esistente. Il punto di partenza per un qualsiasi progetto è la definizione dell'*aspettativa* in funzione della quale si potrà stabilire il successo o meno dello stesso.

Individuati gli obiettivi dell'utente, è necessario creare un insieme di dati *selezionando* un sottoinsieme delle

[1] E' ormai invalso l'uso del termine Data Mining come sinonimo di KDD. In questo lavoro, tuttavia, si preferisce distinguere le due espressioni in accordo a quanto emerso durante la prima conferenza Internazionale su Knowledge Discovery in Database, Montreal, Agosto 1995.

variabili o un campione della Base di Dati da analizzare. L'obiettivo della selezione è identificare le sorgenti di dati disponibili ed estrarne un campione per le analisi preliminari. Oltre le variabili selezionate, è necessario acquisire le corrispondenti informazioni semantiche (*metadati*), indispensabili per capire il significato di ciascuna variabile. I metadati potrebbero includere la definizione dei dati, la descrizione dei tipi, i valori potenziali, il loro sistema sorgente e formato. I dati selezionati devono essere *pre-elaborati* al fine di rimuovere eventuali inconsistenze, rimuovere o ridurre l'effetto del rumore (*noise*), eliminare casi limite (*outlier*), decidere le strategie per la gestione dei valori nulli, scartare dati obsoleti. La pre-elaborazione è quindi finalizzata a migliorare la qualità dei dati selezionati. Spesso potrebbe essere necessario *trasformare i dati* selezionati al fine di isolare caratteristiche utili a rappresentare i dati in base agli obiettivi dell'analisi. In particolare i dati sono trasformati in un formato (modello analitico) compatibile con gli algoritmi disponibili. Il *modello analitico* dei dati rappresenta una ristrutturazione consolidata, integrata e dipendente dal tempo dei dati selezionati e pre-elaborati. Questo passo è fondamentale per garantire l'accuratezza dei risultati che dipende da come gli analisti decidono di strutturare i dati di input.
Una complessa conversione è la *riduzione dei dati*, il cui obiettivo è ridurre il numero totale di variabili da analizzare combinandone alcune tra loro in modo da ottenerne una nuova. Altre possibili conversioni sono lo *scaling* che consente di ridurre input numerici a specifici intervalli, la *discretizzazione* che converte variabili quantitative in categoriche[2], e la *binarizzazione* che converte una variabile categorica in più variabili binarie.
La scelta del task di *Data Mining* (classificazione, regressione, clustering, regole di associazione,ecc..), è una fase fondamentale del processo KDD: essa influenza anche la scelta dell'*algoritmo* per scoprire nuove relazioni, dipendentemente dal tipo di rappresentazione che si decide di usare per le relazioni estratte, e conduce a risultati utili solo se i passi precedenti del processo KDD sono stati correttamente eseguiti.
Infine nel processo KDD le relazioni scoperte devono essere *interpretate* alla luce della conoscenza pregressa, *valutate* rispetto all'obiettivo di business, e comunicate alle parti interessate mediante opportuna reportistica.

## 2. LA NECESSITÀ DI TOOL CHE SUPPORTINO IL PROCESSO KDD

Negli ultimi decenni il ciclo di vita dei processi decisionali nelle aziende è andato accorciandosi sempre più. La tempestività nell'individuare nuovi segmenti di mercato, nello scoprire preferenze e comportamenti da parte dei clienti, nel ridurre eventuali sprechi nella produzione o nel razionalizzare altri processi aziendali, è diventata vitale per la sopravvivenza delle aziende. Tale tempestività, tuttavia, a volte contrasta con la mole dei dati da elaborare per estrarre le informazioni necessarie a supportare il processo decisionale. Il ricorso alle tecnologie dell'informazione è quindi un passo obbligato. E' importante che l'attività di raccolta dei dati non si limiti ai soli dati generati e usati nei processi produttivi o operativi di un'impresa, ma che sia orientata anche ai dati decisionali, caratterizzati da una natura aggregata, una struttura flessibile, un uso non ripetitivo, un orizzonte temporale ampio, e una proprietà di staticità[2].
In un contesto aziendale, la conoscenza scoperta può avere un valore importante perché consente di aumentare i profitti riducendo i costi oppure aumentando le entrate. Questo spiega l'importanza di soluzioni KDD in ambito aziendale.

## 3. TOOL ESISTENTI

Lo studio avanzato nel campo del *Knowledge Discovery*, ha reso possibile lo sviluppo di molti sistemi di Data Mining. Esistono quattro generazioni di Sistemi di Data Mining. I sistemi di prima generazione supportano un piccolo gruppo di algoritmi progettati per valutare vettori di dati. I sistemi di seconda generazione, invece, si interfacciano con basi di dati, per elaborare insiemi di dati complessi e di grandi dimensioni e supportano uno schema di data mining ed un linguaggio di interrogazione, garantendo maggiore flessibilità. I sistemi di terza generazione sono capaci di analizzare dati altamente eterogenei, distribuiti in rete locale e non. I sistemi di quarta generazione, invece, interagiscono direttamente con i generatori di dati.
Sono in commercio diversi sistemi di data Mining tra i quali Clementine[3], WizWhy [4],.DataMind [5].
*DataMind* è un sistema di terza generazione composto da due differenti prodotti che si basano sul modello client-server, rispettivamente *DataMind Professional Editing* e *DataMind DataCruncher*. Il primo, cioè il prodotto lato client, è usato singolarmente o da gruppi di persone per comprendere meglio i dati memorizzati su file locali a cui si connette via ODBC. Il lato server, è un motore per il Data Mining di dati complessi. I modelli creati sono visualizzabili tramite Excel o Word.
*WizWhy* è un sistema di seconda generazione basato su un sofisticato algoritmo matematico per la scoperta di regole di associazione. L'utente deve solo selezionare la base di dati ed il campo variabile dipendente su cui vuole effettuare l'analisi.
*Clementine* è un sistema di seconda generazione che permette all'utente finale di scoprire informazioni, interessanti ed utili, nascoste all'interno di grandi basi di dati. Esso è dotato di un'interfaccia visuale che assiste l'utente nelle fasi del processo KDD.

## 4. KDB2000

KDB2000 (www.di.uniba.it/~malerba/software/KDB2000) è un sistema di seconda generazione che supporta in

[2] Per *variabili categoriche* i possibili valori sono finiti e diversi e si distinguono in due sottotipi: *variabili nominali* e *ordinali*. Le prime esprimono il nome del tipo di oggetto a cui si riferiscono, ma non prevedono alcun ordinamento tra i valori distinti. Le variabili ordinali, invece, prevedono un ordinamento tra i possibili valori distinti. Anche le variabili *quantitative* si distinguono in due sottotipi: *continue* e *discrete***.** Le prime hanno valori reali, le seconde, invece, prevedono valori interi.

maniera intelligente l'utente nelle diverse fasi del processo KDD, allo scopo di estrarre e interpretare opportunamente relazioni tra i dati analizzati.

### 4.1 Architettura del Sistema

L'architettura funzionale di KDB2000 è mostrata in Figura 1.
Il Data Banker è quella componente contenente i driver verso i dati disponibili in una base di dati accessibile tramite fonte dati ODBC. La componente Data Banker gestisce la creazione, l'accesso e la modifica dei dati.
La *Data Visualization*(DV) è quella componente che consente di visualizzare i risultati di ciascuna fase KDD. Gli strumenti *ETL* (*Extraction, Transformation* and *Loading*) servono ad estrarre i dati dalla base di dati, a pulirli in modo da renderli consistenti con lo schema, a riepilogarli e a convertirli nel formato compatibile con quello previsto dai metodi di analisi disponibili. La componente di *Data Management* gestisce la creazione, l'accesso e la visualizzazione dei dati. Tale componente gestisce, inoltre, i *metadati*, vale a dire le strutture informative che descrivono le sorgenti dei dati. La componente *Data Mining tools*, invece, ospita gli algoritmi di Data Mining per l'analisi dei dati selezionati.

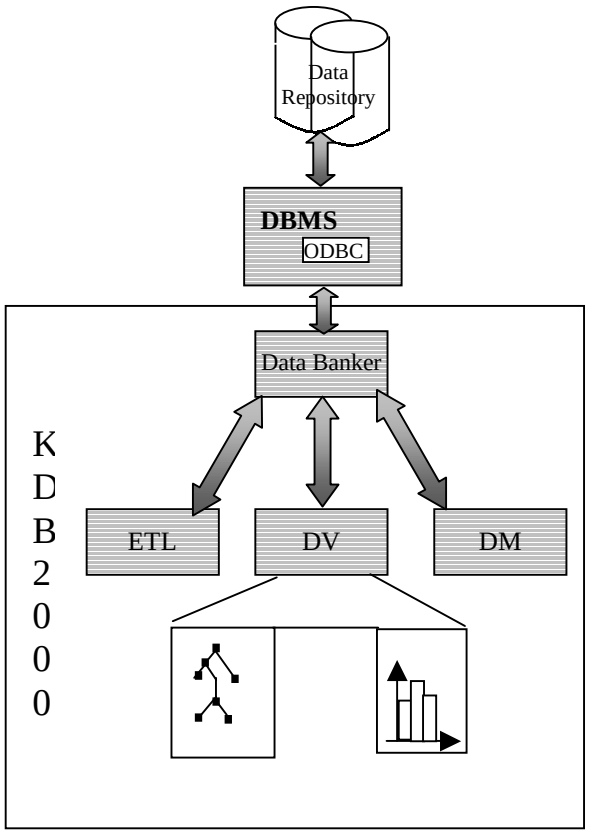


Figura 1 Architettura di KDB2000

### 4.2 Funzionalità disponibili nel Sistema

KDB2000 mette a disposizione dell'utente un insieme di funzionalità che lo assistano nelle fasi del processo KDD dall'acquisizione dei dati alla visualizzazione dei risultati. Nella fase d'acquisizione dei dati, l'utente può selezionare o creare una nuova fonte dati per accedere e gestire le informazioni contenute nella base di dati relativa. L'obiettivo della selezione è identificare le sorgenti di dati disponibili ed estrarre i dati di interesse. I dati possono essere selezionati mediante un'interrogazione SQL esplicitamente definita dall'utente o attraverso l'ausilio di un wizard di creazione della query: tutte le interrogazioni possono essere memorizzate con i relativi risultati per utilizzi futuri in una workspace utente. La fase di pre-eleborazione dei dati utilizza strumenti sia statistici sia di visualizzazione grafica per aumentare la qualità degli stessi. In particolare per dati quantitativi possono essere calcolate le seguenti statistiche: il massimo, il minimo, la media, la moda, la mediana, la deviazione standard e la correlazione. Le tecniche di visualizzazione grafica, invece, permettono una migliore comprensione del contenuto dei dati tramite istogrammi, diagrammi a torta e scatter plot. La fase di trasformazione dei dati converte i dati esistenti in un formato più adeguato alle esigenze dell'utente e compatibile con gli algoritmi di Data Mining . Le funzionalità implementate sono:

- *discretizzazione*: permette, a seconda del numero di intervalli inserito, di partizionare l'insieme dei valori (quantitativi), ottenendo variabili categoriche;
- *scaling*: lavora su dati quantitativi e permette il passaggio da un intervallo di riferimento attuale ad un intervallo di riferimento opportunamente scelto dall'utente;
- *sostituzione dei valori nulli*: permette di sostituire i valori nulli presenti nella base di dati con la media o il valore più frequente a seconda che si tratti di variabili quantitative o categoriche.
- *campionamento*: restituisce un sottoinsieme casuale dell'insieme di partenza, di dimensione pari ad una percentuale fissata.

Le fasi appena descritte sono preparatorie al task di Data Mining che lavora su domini resi disponibili proprio dall'analisi dei dati. I metodi di Data Mining implementati sono:

- *Alberi di decisione*: ricerca una funzione che mappi opportunamente una collezione di dati in classi diverse e predefinite [6];
- *K-nearest neighbor*: classifica un'osservazione in base alle k osservazioni più vicine secondo una misura di distanza fissata a priori [7];
- *Alberi di modelli*: ricerca una funzione che mappi una collezione di dati in un valore reale predetto per una data variabile [8];
- *Clustering*: cerca di identificare un insieme finito di categorie o raggruppamenti (cluster) per descrivere i dati [9];
- *Regole di associazione*: ricerca una descrizione compatta di un sottoinsieme di dati. In particolare si generano regole del tipo
  X->Y dove X e Y sono insiemi di caratteristiche [10][11].

La valutazione dell'accuratezza predittiva degli alberi di decisione e alberi di modelli avviene per mezzo della strategia *k-fold Cross-Validation*. (k-CV). L'applicazione del k-CV, prevede una suddivisione casuale dell'insieme di training in k sotto insiemi disgiunti $D_1$, …, $D_k$ ciascuno contenente

approssimativamente lo stesso numero di osservazioni. Si costruisce il modello di Data Mining per ciascun insieme $D_i$, i= 1…k, usando D-$D_i$ come insieme di addestramento: questo modello è poi testato sull'insieme $D_i$. Lo stesso processo è ripetuto per i k sottoinsiemi.

### 4.3 Ambiente di Sviluppo e piattaforma

KDB2000 è sviluppato in Microsoft Visual C++ 6.0, un ambiente complesso e potente che consente di realizzare applicazioni a 32 bit per Windows 95/98/NT. I requisiti minimi per l'installazione di KDB2000 sono: spazio su disco 100 MB, memoria centrale 64MB. Il fabbisogno di memoria centrale da parte di KDB2000 è proporzionato alla mole dei dati da analizzare poiché alcuni metodi di Data Mining caricano in memoria centrale i dati da elaborare.

## 5. UN ESEMPIO COMPLETO DI COME SI PUÒ ESTRARRE CONOSCENZA CON KDB2000.

KDB2000 è un valido strumento per l'estrazione di conoscenza, utilizzabile anche nell'ambito scientifico. Un esempio potrebbe essere uno studio scientifico condotto da un gruppo di biologi, allo scopo di trovare un predittore affidabile per l'età degli abalone (organismi marini dotati di guscio).
Il primo passo è selezionare l'insieme di caratteristiche da analizzare in relazione all'obiettivo proposto. A tal scopo è utile valutare caratteristiche fisiche come: sesso, lunghezza del corpo, diametro, altezza, numero di anelli, peso totale, peso dell'apparato scheletrico, del guscio e degli organi interni. Ciascun attributo è pre-elaborato tramite tecniche statistiche come: l'analisi dei minimi e massimi, lo studio di deviazione standard, la ricerca di valori nulli…, allo scopo di valutare la presenza di outlier o incertezza nei dati. La visualizzazione grafica tramite diagrammi a torta, istogrammi o scatter-plot è, inoltre, un valido strumento per rappresentare la distribuzione dei valori. Tali pre-elaborazioni suggeriscono come trasformare i dati selezionati. Per esempio, si può decidere di rimuovere la presenza di valori nulli, sostituendoli con la media o la moda dell'attributo a seconda che esso sia continuo o discreto; oppure di discretizzare o ridurre (scaling) un attributo numerico, o di estrarre dai dati un campione casuale, analizzabile con algoritmi di Data Mining con migliori prestazioni in tempo. L'algoritmo di Data Mining, disponibile in KDB2000 per la scoperta di un predittore dell'età degli abalone, è *Smoti* (*S*tepwise *Mo*del *T*ree *I*nduction)[8] che costruisce un albero di modelli. Il modello costruito è in forma di albero (Figura 2) o di insieme di regole.

## Ringraziamenti

Questo lavoro è parte del progetto MURST COFIN-1999 progetto sul " *Modelli Statistici di Classificazione e di Segmentazione per l'Analisi di Dati Strutturati in Forma Complessa: Metodologie, Software e Applicazioni*"". Gli autori ringraziano Marcello Lucente per aver collaborato alla realizzazione di KDB2000.

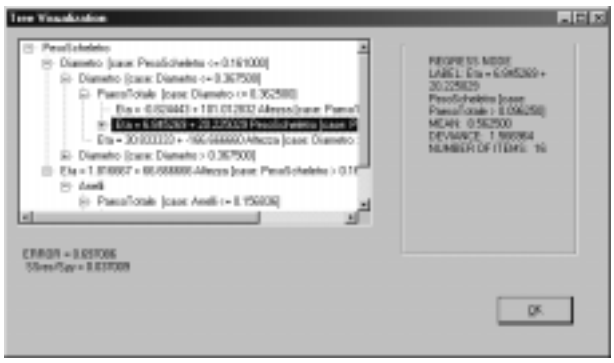


Figura 2 KDB2000: Visualizzazione albero di modelli

## Bibliografia

[1] Fayyad U., Piatesky-Shapiro G., Smyth P. (1996). *From Data Mining to Knowledge Discovery: an Overview*, Advances in Knowledge Discovery and Data Mining, pp. 1-34, AAAI Press.

[2] Immon, W.H.(1992). *Building the data warehouse*. Wellesley, MA:QED Tech. Pub, Group.

[3] http://www.isl.co.uk/elem.html

[4] http://www.wizsoft.com/why.html

[5] http://www.datamindcorp.com

[6] Quinlan, J. R., (1993), *C4.5: Programs for Machine Learning*, San Mateo, Calif.: Morgan Kaufmann.

[7] Mitchell, (1997). *Machine Learning*, McGraw Learning.

[8] D. Malerba, A. Appice, A. Bellino, M. Ceci, & D. Pallotta (2001). Stepwise Induction of Model Trees. In F. Esposito (Ed.), *AI*IA 2001: Advances in Artificial Intelligence*, Lecture Notes in Artificial Intelligence, 2175, Springer, Berlin, Germany.

[9] Farnstrom, F, Lewis, J, Elkan, C (2000). Scalability for Clustering Algorithms Revisited. *SIKDD Explorations*,Vol. 2, n. 1, pp. 51-57.

[10] Agrawal, R, Imielinski, T, and Swami, A. Mining association rules between sets of items in large databases. *Proc. Of the ACM SIGMOD Conference on Management of Data*, Washington, D.C., May 1993.

[11] Agrawal, R., Srikant, R.(1994). Fast Algorithms for Mining Association Rules. *Proc. of the Twentieth VLDB Conference*, Santiago: Cile .