DIPARTIMENTO DI ELETTRONICA, INFORMATICA
E SISTEMISTICA

Università di Bologna
Area Progetti Integrati Cesena (APICe)

Introduzione al Data Mining:
Potenzialità, Applicazioni e Progetti di Ricerca

Seminario per il corso di laurea specialistica in
Ingegneria Informatica (II Facoltà)

Gianluca Moro
DEIS - Università di Bologna
Via Venezia 52 - 47023 Cesena
gianluca.moro@unibo.it

Archivi di dati = "Tombe" di dati ?
La necessità è la madre delle invenzioni

- La capacità di raccogliere e memorizzare dati ha largamente superato la capacità umana di analizzarli
  - Strumenti di raccolta automatica dei dati, maturità della tecnologia database
  - → Enormi quantità di dati memorizzati e disponibili

    distanza crescente fra la *generazione* dei dati e la loro *comprensione*
- Siamo assetati di conoscenza ma anneghiamo nei dati
- Tuttavia, i dati contengono informazioni di grande interesse economico, sociale e scientifico:
  - la ricerca nel Data Mining ha come scopo la progettazione di strumenti per trasformare i dati in informazione
- Data warehousing e data mining
  - Integrazione, analisi/sintesi ed estrazione di conoscenza

A quali domande risponde il data mining

- Da cosa è influenzata la vendita di un certo prodotto ?
  - Market Basket Analysis
- In quali macro gruppi si suddividono i miei clienti ?
  - Segmentazione dei clienti
- Quale prodotto devo proporre ad un dato cliente ?
  - Cross-selling
- Quali sono i clienti che potrei perdere ?
  - Customer retention
- Quale altro prodotto devo proporre all'atto di un acquisto ?
  - Up-selling
- Qual è il rischio che corro a fronte di un investimento ?
- Quali sono le correlazioni tra i fenomeni che caratterizzano la mia realtà aziendale ?

Differenze nell'analisi tra Data Warehousing e Data Mining

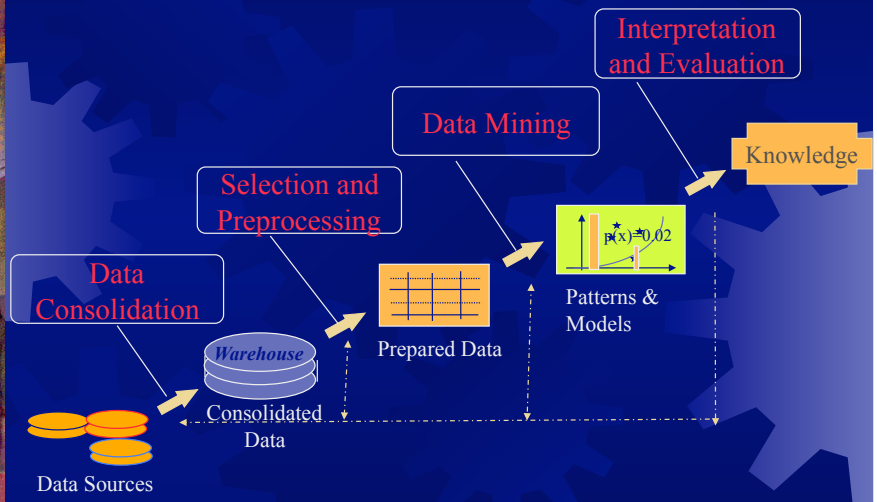
- Data Warehousing:
  - *Vorrei visualizzare le vendite di ogni prodotto suddivise per punto vendita*
- Data Mining:
  - *Vorrei sapere quali sono le caratteristiche dei punti vendita con redditività alta*
- ***Nel primo caso l'utente sa già quello che cerca***
- ***Nel secondo caso vuole scoprire la causa di un effetto***

Definizione di *knowledge discovery*

The nontrivial extraction of implicit, previously unknown and potentially useful information from data

W. Frawley, G. Piatetsky-Shapiro, and C. Matheus: "Knowledge Discovery in Databases: An Overview". AI Magazine, Fall 1992, pgs 213-228

Il processo di Knowledge Discovery



Algoritmi e Tecniche di Data Mining (i)

- Partono da un insieme di **osservazioni**:
  - Studenti, prodotti, clienti, pazienti, ...
  - Un'osservazione è caratterizzata da un insieme di attributi
  - Studente: [Facoltà, numero esami sostenuti, media voti]
  - Paziente: [Età, sesso, risultato di un certo esame clinico]

Att1	Att2	Att3	Att4	Att5
0.54	Giallo	Y	123	…
0.89	Rosso	N	5734	…
2	Verde	N	8944	…

- Scoprono informazioni relative alle osservazioni in input
  - Correlazioni tra valori in colonne diverse
  - Correlazioni tra osservazioni/righe diverse

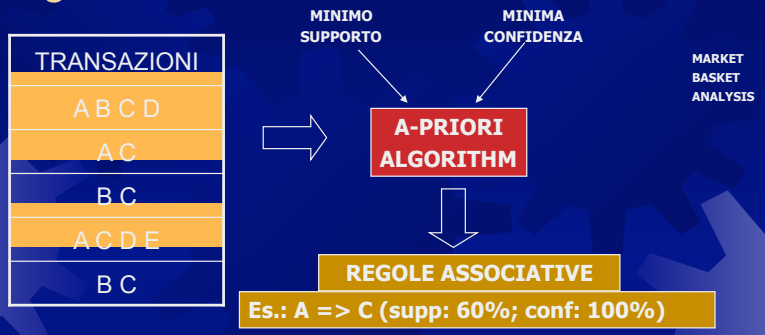
Algoritmi e Tecniche di Data Mining (ii)

UNSUPERVISED LEARNING

	Esempi
• REGOLE ASSOCIATIVE ▪ Scopre le correlazioni tra colonne.	Utilizzato per analizzare gli acquisti in un supermercato
• CLUSTERING ▪ Scopre i cluster (raggruppamenti) di osservazioni/righe simili tra di loro	Utilizzato per segmentare i clienti
• CLASSIFICAZIONE ▪ Predice il gruppo di appartenenza di un'osservazione	Per riconoscere i clienti che stanno per abbandonare l'azienda
• REGRESSIONE ▪ Stima il valore di un attributo numerico di un'osservazione	Fare stime: es. Quanto sarà l'incasso di oggi, nota la data e le condizioni meteo ?
• SERIE TEMPORALI ▪ Prevede l'andamento di un certo valore	

SUPERVISED LEARNING

Regole Associative



- SUPPORTO DI UNA REGOLA: percentuale di transazioni che contengono sia antecedente che conseguente (es: |A U C| / T)
- CONFIDENZA DI UNA REGOLA: perc. delle transazioni contenenti l'antecedente, che contengono anche il conseguente
  |A U C | / | C |
- OSSERVAZIONE: Dato un ItemSet non frequente (< minSupp), i suoi sovrainsiemi sono non frequenti. -> questo rende l'algoritmo più veloce

Regole Associative (i)

- Transazione
  - insieme di elementi (item) acquistati congiuntamente (quello che si trova in un carrello della spesa)
- Regola Associativa
  - dato un insieme di item I e un insieme di transazioni D, una regola associativa del tipo $X \Rightarrow Y$ (X *implica* Y) (con $X,Y \subset I$ e $X \cap Y = \varnothing$) è un'implicazione

chi compra X compra anche Y

Regole Associative: Supporto e Confidenza

- Supporto di una regola (| X U Y | / | T |)
  - È la percentuale di transazioni che contengono sia X che Y sul totale delle transazioni esistenti

    es: (il 40% delle transazioni natalizie include panettone e champagne)
- Confidenza di una regola (| X U Y| / |X|)
  - È la percentuale di transazioni che contengono sia X che Y rispetto alle transazioni che contengono almeno X

    es: (a Natale, l'80% delle persone che comprano champagne comprano anche il panettone)
- Problema:
  - determinare tutte le regole associative che abbiano supporto almeno pari a MINSUPP e confidenza almeno pari a MINCONF

Supporto e confidenza: Esempio

- La regola A $\Rightarrow$ C ha
  - Supporto pari al 50%, perché {A C} compare in 2 transazioni su 4
  - Confidenza pari al 66%, perché su 3 transazioni in cui compare A, in due compare anche C
- La regola C $\Rightarrow$ A ha
  - Supporto pari al 50%
  - Confidenza pari al 100%

Transaction ID	*Items*
100	A B C
200	A C
300	A D
400	B E F

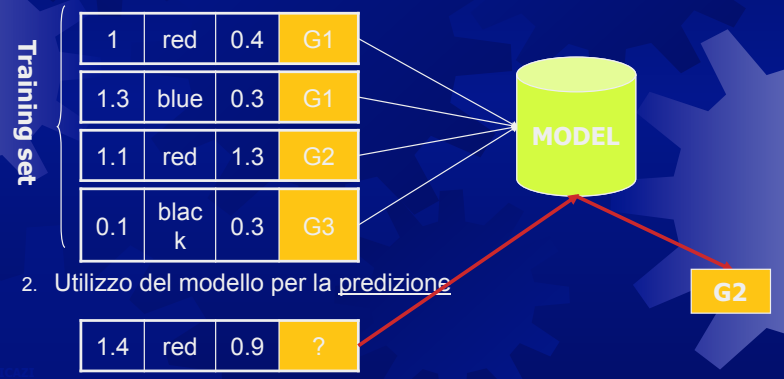
Regole Associative: Utilità

- Trova le regole che hanno noccioline nel conseguente
  - *possono essere usate per capire quali prodotti il supermercato deve comprare per favorire la vendita di noccioline*
- Trova le regole che hanno noccioline nell'antecedente
  - *può prevedere quali prodotti possono subire una riduzione delle vendite se il supermercato decide di non vendere più noccioline*
- Trova le regole che hanno noccioline nell'antecedente e birra nel conseguente
  - *può servire per capire quali altri prodotti oltre alle noccioline servono per favorire la vendita di birra*
- Trova le regole che riguardano item delle corsie 10 e 11
  - *possono essere usate ai fini di una migliore organizzazione dei prodotti nelle corsie*
- Trova le regole più interessanti
  - *Ad esempio regole con maggiore confidenza e/o supporto*

Classificazione

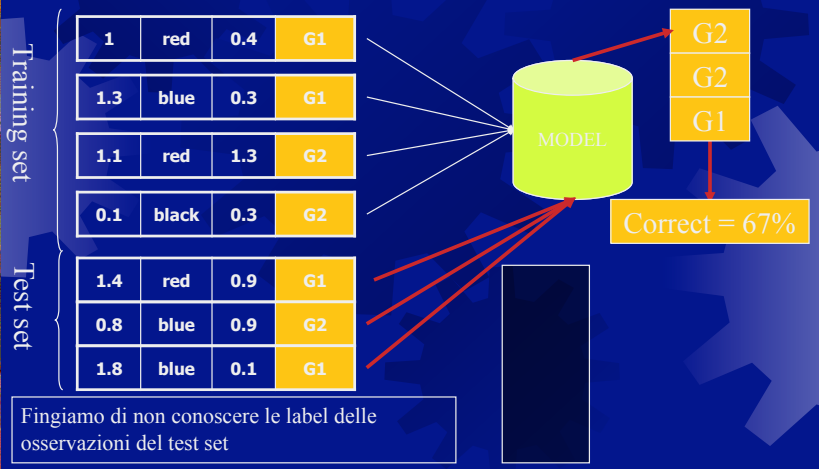
- SCOPO: Date delle osservazioni etichettate (con un attributo categorico), addestrare una macchina ad etichettare nuove osservazioni (in cui l'attributo non è noto)

1. Training del modello



2. Utilizzo del modello per la predizione

Classificazione: Misurare l'errore (hold out)



Applicazioni e Tecniche

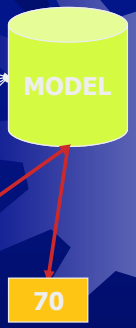
- Applicazioni:
  - Capire se una certa persona è un buon cliente o no
  - Diagnosi di malattie
  - Individuazione delle frodi
  - Targeted Mailing
- Tecniche:
  - Decision Trees, Neural Networks, Baysian Networks

Regressione

- SCOPO: Dato un insieme di osservazioni e un attributo numerico *target*, addestrare una macchina ad indovinare l'attributo *target* di una nuova osservazione, di cui si conoscono tutti gli altri attributi

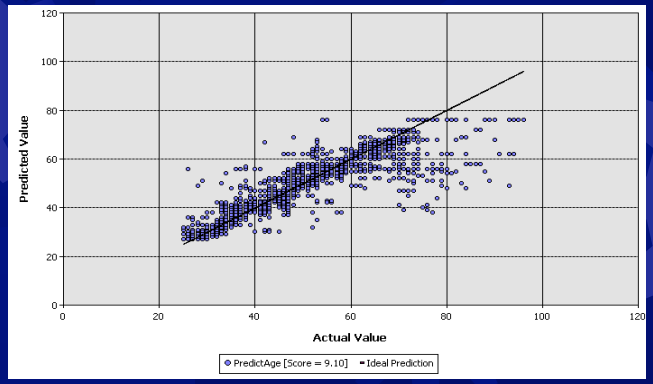
Training set

ALTEZZA	PESO	SESSO
165	60	M
167	52	F
179	80	M
181	61	F
172	?	M



Regression Decision Trees

- Idea:
  - Il funzionamento è quello dei Decision Tree, con la differenza che ad ogni split si cerca di minimizzare la varianza all'interno di ciascun nodo figlio



Applicazioni e Tecniche

REGRESSIONE

- Applicazioni:
  - Stimare il ricavo che un cliente genererà
  - Stimare la probabilità che una persona risponda ad un annuncio pubblicitario
  - *Classificare* (avendo come output un valore numerico anziché una label)
- Tecniche:
  - Regression Decision Trees
  - Regressione Lineare (Singola / Multivariata)
  - Regressione Non Lineare

Problemi tipici di Data Mining: Le Mucche della nuova Zelanda

- Ogni anno, gli allevatori caseari in Nuova Zelanda devono prendere una difficile decisione: quali capi tenere nell'allevamento e quali vendere per la macellazione
- Un quinto dei capi degli allevamenti è abbattuto ogni anno alla fine della stagione del latte, quando il foraggio inizia a scarseggiare
- La storia di produzione di vitelli e di latte di ogni bovino influenza questa decisione, insieme ad età, salute, storia, comportamento etc.
- Sono stati registrati circa 700 attributi per milioni di capi
- *Come estrarre da questi dati la conoscenza implicita nelle decisioni degli allevatori di maggior successo ?*

Problemi tipici di Data Mining: Gli agricoltori statunitensi

- il dipartimento dell'agricoltura degli Stati Uniti ogni anno eroga indennizzi per danni da maltempo a centinaia di migliaia di agricoltori
- una frazione delle richieste di indennizzo è fraudolenta
- un'analisi a campione delle richieste per verificarne l'autenticità ha un costo molto elevato rispetto alla resa
- un progetto di data mining volto a individuare le frodi ha reso oltre venti volte il suo costo

Problemi tipici di Data Mining: Università (negli USA)

- Ogni anno le Università statunitensi ricevono domande di ammissione da parte di studenti
- Gli studenti forniscono una serie di dati sulla carriera scolastica e personali
- Obiettivo: scegliere gli studenti migliori, ossia che completeranno gli studi con ottimi voti e senza ritardi
- Dalla storia degli studenti che hanno già completato gli studi individuare gli studenti migliori e le loro peculiarità
- *A partire da queste caratteristiche è possibile stimare per ogni studente candidato voto finale e ritardo alla laurea ?*
- Progetto simile col consorzio AlmaLaurea
  - Stima della fascia di voto e ritardo per gli studenti che dalla laurea triennale si iscrivono alla specialistica

Progetti e Ricerca

Gruppo Informatica

SiMET
Sistema di Monitoraggio dell'Economia del Territorio

Referente DEIS: Gianluca Moro

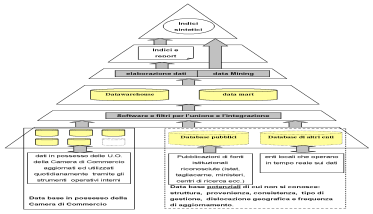
Camera di Commercio Industria Artigianato Agricoltura di Forlì-Cesena

Dipartimento di Elettronica Informatica e Sistemistica

CISE Centro per l'Innovazione e lo Sviluppo Economico – Azienda Speciale della Camera di Commercio di Forlì-Cesena

Il progetto

- nasce dall'esigenza di **rilevare e misurare** le principali **dinamiche economiche** del territorio, attraverso l'ampia mole di informazioni e **dati** provenienti **da fonti locali**, nazionali ed internazionali, con l'impiego di **tecniche innovative di data mining.**
- **SiMET** mira ad offrire un **sistema di indicatori** economici che permettano di **monitorare** e **valutare** le **variabili strutturali e congiunturali** dell'economia locale attraverso una accurata scomposizione dei fattori chiave.



Obiettivi

- Individuazione, digitalizzazione, aggiornamento e verifica delle basi dati della Camera di Commercio.
- Individuazione/*Scoperta* di fenomeni rilevanti e definizione dei relativi panel di indicatori.
- Condivisione del sistema con gli altri Enti Istituzionali del Territorio per ampliare le basi dati di riferimento e individuare azioni locali ad ampio raggio.

Metodologie, Modelli e Tecnologie

Progettazione e realizzazione del data warehouse:

processi di importazione e cleaning dei dati, modellazione multidimensionale, set di datamart, *definizione di algoritmi di mantenimento incrementali a consistenza completa* etc.

Progettazione e sviluppo di sistemi web per la navigazione e l'elaborazione degli indicatori.

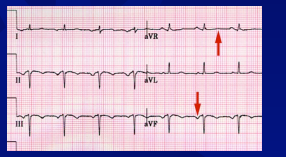
Definizione di algoritmi di data clustering per il sistema di data mining & knowledge discovery

ALMA MATER STUDIORUM UNIVERSITA' DI BOLOGNA
DEIS, SEDE DI CESENA

Progetto: diagnosi automatica del carcinoma al seno da dati di risonanza magnetica

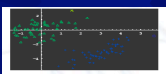
- in collaborazione con Dipartimento Interaziendale di Oncologia (AUSL Forlì)
- 23% di tutti i tumori femminili ed è in aumento, 12% delle donne in U.S. (2002 - National Center for Health Statistics)
- tasso di sopravvivenza del 73% (regioni sviluppate)
- recentemente la risonanza in senologia ha mostrato accuratezza diagnostica superiore alla mammografia ed ecografia (68% - 79%)
- Sviluppo di nuove metodologie di diagnosi mediante tecniche di data mining innovative:
  - che sfruttino l'informazione contenuta nelle immagini 3D
  - che forniscano misure dell'affidabilità della previsione

Progetto: Predizione di Battiti Cardiaci Pre-Ectopici



- Battiti extrasistoli (ectopici) possono causare fibrillazione
  - Possibili cause: aritmie cardiache, emodialisi
- piccola scarica elettrica poco prima (max 3-4 battiti) di un battito ectopico previene il battito ectopico stesso
- Problema:
  - i battiti pre-ectopici apparentemente sono battiti normali con forme d'onda usuali
  - Come distinguere questi battiti da quelli non pre-ectopici?
- Primi risultati confortanti
  - Data set: 10 pazienti in emodialisi, 97 attributi per ogni battito, 1 ora di 10 tracciati contiene 97*80*60*10 ≈ 5.000.000 valori
  - Idea: etichettare i battito con la distanza dal battito ectopico e determinare con la regressione quali dei 97 attributi stimano meglio questa distanza
  - Bastano 25 attributi: falsi negativi 2%, falsi positivi 8%

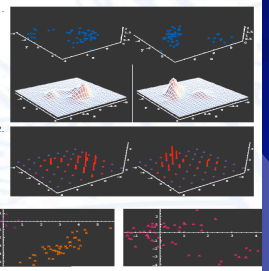
Data Clustering Distribuito



- *Data clustering* partizionamento dei dati che massimizza la similarità intra-gruppo e la dissimilarità inter-gruppo
- In ambiente distribuito, spesso il trasferimento dei dati distribuiti verso un unico sito non è realizzabile (dati sensibili, costo di trasmissione elevato, vincoli temporali)

Lo schema KDEC per il data clustering

1. Stima statistica della densità calcolata localmente a ciascun sito
2. Campionamento regolare
3. Trasmissione dei campioni a un sito ausiliario e somma dei campioni omologhi
4. Trasmissione della somma a ciascun sito e interpolazione locale
5. Clustering locale utilizzando la stima globale interpolata
6. Lo schema può essere implementato in architetture client/server, multiagente e peer-to-peer

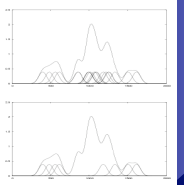


Possibili Applicazioni

- Estrazione automatica di nuova conoscenza da Sistemi Informativi Distribuiti su diverse sedi di una medesima organizzazione, ad esempio:

  ricavare dati sulla salute pubblica a partire da dati del sistema sanitario depositati nei singoli Ospedali locali e nazionali senza concentrare i dati stessi su un'unica macchina

  derivare informazioni di sintesi sullo stato economico di un intero territorio analizzando i dati locali ad Istituti Bancari o non commerciali come le camere di commercio, senza spostare i dati stessi dai singoli gestori, preservando perciò privacy e sicurezza dell'informazione



Collaborazioni

- Deutsche Forschungszentrum fuer Kuenstliche Intelligenz, Saarbruecken, Germany
- Università di Firenze

Progetti

- "Il cittadino europeo nell'e-governance: profili filosofico-giuridici, giuridici, informatici ed economici" (PRIN 2003)

Stream Data Clustering

- Obiettivo
  - clusterizzare un flusso di dati multi-dimensionale (eventualmente infinito) garantendo una soglia max al num. di operazioni
- Velocità vs accuratezza
  - rendere trascurabile il num. di errori
- DeStream: Idea di base
  - Il dato/punto cade in una griglia multi-dimensionale
  - ricalcolo della densità dei punti della griglia all'interno di un'iper-sfera di raggio parametrizzabile in base alle prestazioni richieste
  - aggiornamento del clustering all'interno dell'ipersfera
- Applicazioni:
  - stream delle transazioni con carte di credito
  - sistemi di monitoraggio evoluti

Un paio di nuovi casi aziendali

- Consorzio Agrario
  - 30 agenzie distribuite sul territorio romagnolo, vendita di concimi, fitofarmaci, sementi, ma anche carburanti e attrezzature
  - Oltre 500000 record all'anno
  - Progetto: valutare la fedeltà dei clienti
- AWS
  - il principale corriere espresso a capitale tutto italiano
  - Oltre 120 terminal in tutta italia
  - Milioni di record all'anno
  - Progetto:
    - stimare il tempo di consegna reale man mano che il pacco procede verso la destinazione, valutando anche possibili percorsi alternativi da quello pianificato che offrano maggiori probabilità di ridurre il tempo di consegna
    - Individuare indicatori di qualità dei terminal in grado di prevedere con un certo anticipo anomalie e problemi locali ai terminal stessi

Alcune pubblicazioni sui temi trattati

- Data Mining
  - S. Lodi, G. Moro, C. Sartori *Stream Clustering Based on Kernel Density Estimation.* In The 17th European Conference on Artificial Intelligence, Riva del Garda, Italy, 2006.
  - J. Costa da Silva, M. Klusch, S. Lodi and G. Moro, *Privacy-preserving agent-based distributed data clustering.* Web Intelligence and Agent Systems: An international journal, Vol. 4, No. 3, pp. 1-18, IOS Press, 2006.
  - J. Costa Da Silva, M. Klusch, S. Lodi, G. Moro, *Inference attacks in peer-to-peer homogeneous distributed data mining.* In the 16th European Conference on Artificial Intelligence (ECAI 2004), IOS Press, Vol. 110, Valencia, Spain, 2004.
  - M. Klusch, S. Lodi, G. Moro. *The Role of Agents in Distributed Data Mining.* In the 2003 International Conference on Intelligent Agent Technology (IAT'2003), IEEE Computer Society press, Halifax, Canada 2003.
  - M. Klusch, S. Lodi, G. Moro. *Distributed Clustering Based on Sampling Local Density Estimates.* In the Eighteenth biennial International Joint Conference on Artificial Intelligence (IJCAI'2003), Morgan Kaufmann, Mexico, 2003. (18% acceptance rate)
  - M. Klusch, S. Lodi, G. Moro. *Issue on agent-based distributed data mining.* In the Second International Joint Conference on Autonomous Agents and Multi Agent Systems (AAMAS'2003) ACM, Melbourne, Australia July 2003.
- Reti di sensori
  - G. Monti, G. Moro, S. Lodi: *W*-Grid: A Robust Decentralized Cross-layer Infrastructure for Routing and Multi-Dimensional Data Management in Wireless Ad-Hoc Sensor Networks,* In the 7th IEEE International Conference on Peer-to-Peer Computing (P2P 2007), Galway, Ireland, 2007 (18% acceptance rate)
  - G. Moro, G. Monti *W-Grid: a Cross-Layer Infrastructure for Multi-Dimensional Indexing, Querying and Routing in Ad-Hoc and Sensor Networks.* In the Sixth IEEE International Conference on Peer-to-Peer Computing (P2P 2006), Cambridge, UK, 2006. (19% acceptance rate)
  - G. Moro, G. Monti and A.M. Ouksel, *Routing and Localization Services in Self-Organizing Wireless Ad-Hoc and Sensor Networks Using Virtual Coordinates.* In the IEEE International Conference on Pervasive Services 2006 (ICPS 2006), Lyon, France, 2006.
  - G. Monti, G. Moro, C. Sartori: W^R-Grid: A Scalable Cross-Layer Infrastructure for Routing, Multi-dimensional Data Management and Replication in Wireless Sensor Networks. LNCS 4331 Springer 2006

Tecnologie per il Data Mining

- Open Source
  - Weka (The University of Waikato, New Zealand)
    - http://www.cs.waikato.ac.nz/ml/weka/
  - RapidMiner (Ingo Mierswa et. al.)
    - http://rapid-i.com/index.php?lang=en
  - R (The R Foundation for Statistical Computing)
    - http://www.r-project.org
  - ADaMSoft (Consorzio inter-universitario CASPUR)
    - http://adamsoft.caspur.it/
- Commerciali
  - SQL Server 2005 Analysis Service
  - Oracle Data Mining
  - IBM DB2 Intelligent Miner
  - …..

Altri riferimenti

- Testi
  - Jiawei Han e Micheline Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers, 2000. ISBN 1-55860-489-8
  - Data Mining: Practical Machine Learning Tools and Techniques (Second Edition) Ian H. Witten, Eibe Frank, Morgan Kaufmann, 2005, ISBN 0-12-088407-0 (riferimenti a Weka)
- Database disponibili online
  - http://mlearn.ics.uci.edu/MLSummary.html (University of California at Irvine)