Data Mining Opportunities in Engineering: What, Why and How?

Paulo Cortez

Department of Information Systems (DSI)/Algoritmi R&D Center

University of Minho

www.dsi.uminho.pt/~pcortez

pcortez@dsi.uminho.pt

Business Intelligence Group (BIG), DSI, U.Minho.

- Develops teaching and R&D activities in: Artificial Intelligence, Data Mining...
- Successful applications: Database Marketing, Corporate Bankruptcy Prediction , Water Dam Quality, Civil Engineering, …
- Recent Projects:
  - Grid Data Mining (FCT);
  - Internet Congestion Control Using Neural Networks (CRUP/British Council);

What is Data Mining?

- **Data Mining** is a new field (since 1996) that derived from the disciplines of **Databases**, **Artificial Intelligence** and **Statistics**;
- **Data Mining** is also known as **Knowledge Discovery from Databases (KDD)**;
- **Data Mining/KDD** is the process of extracting **useful knowledge** from raw data;
- **Data Mining** includes several **iterative and interactive steps**: domain understanding, data selection, preprocessing and transformation, application of algorithms to find patterns, validation and interpretation and use of knowledge [Fayyad et al, 1996].

What is Data Mining?

Data Mining goals: Prediction

Regression: estimate a numeric (dependent) output value from several (independent) input variables

Algorithms: Linear Regression, Neural Networks (MLP, RBF,…), Support Vector Machines, PLS, Regression Tree, Random Forest, K-Nearest Neighbor, MARS, BRUTO, …

Opportunity: Very often, engineering applications can be defined in terms of regression tasks!!!
E.g. Predict the rise time of a robot arm; estimate the fuel consumption of a vehicle; predict the resistance of steel beams, …

What is Data Mining?

Data Mining goals: Prediction

Classification: label (output) an item given some of its characteristics (input variables)

Algorithms: Linear Discriminant Analysis, Naïve Bayes, Neural Networks (MLP, RBF,…), Support Vector Machines, Decision Tree, Random Forest, K-Nearest Neighbor, …

Opportunity: Classification is the most used Data Mining task! E.g. What is the type of soil (grey, vegetation, …) that corresponds to a sattelite (landsat) image? Discriminate sonar signals bounced off a metal cylinder (class "M") and a roughly cylindrical rock ("R"); Classify a given building according to its response to earthquakes (“bad“, “medium”, “good”), …

What is Data Mining?

Data Mining goals: Description

Clustering: segmentation of data into clusters with similar characteristics

Algorithms: Kohonen NNs (SOM), EM, K-Means, …

Opportunity: When no labels (outputs) are defined, clustering can be used to create the output classes used for classification.

E.g. Definition of a new classification index for a given engineering area; Identifying groups of houses according to their value, geographical location; …

What is Data Mining?

Data Mining goals: Description

- **Summarization:** get a compact description of the data
  - Techniques: statistics (e.g. mean, std), sumarization rules, visualization algorithms, …
  - E.g. Show the monthly shoe sales, …
- **Association Rules:** used on transactional data
  - Algorithms: Apriori
  - E.g. Market-basket analysis (“64% of the clients that bought milk also purchased bread”), …

What is Data Mining?

Some of my own application examples:

- Prediction of Internet traffic in the UK research and academic network (UKERNA);
- Mortality assessment in Intensive Care Units;
- Lamb meat quality (tenderness) assessment;
- Forest fire area prediction using meteorological data;
- Security: intrusion detection using video images;

Why Data Mining?

- Due to the advances in Information and Communication Technologies (ICT), it is **easy** to collect and **store** data;
- **Vast databases** are available (the amount of **stored data doubles** every 9 months!);
- All this data **holds valuable information**;
- Human **experts are limited** and may overlook important details;
- Classical **statistical analysis** (e.g. multiple regression) **breaks down** when such **vast amount** and/or **complex data** is present.

The alternative is to use **(semi)automated** discovery tools!

Opportunity: most engineering problems are nonlinear, thus nonlinear DM methods (e.g. NNs, SVMs, PLS, Decision Trees, Random Forest) should work well!

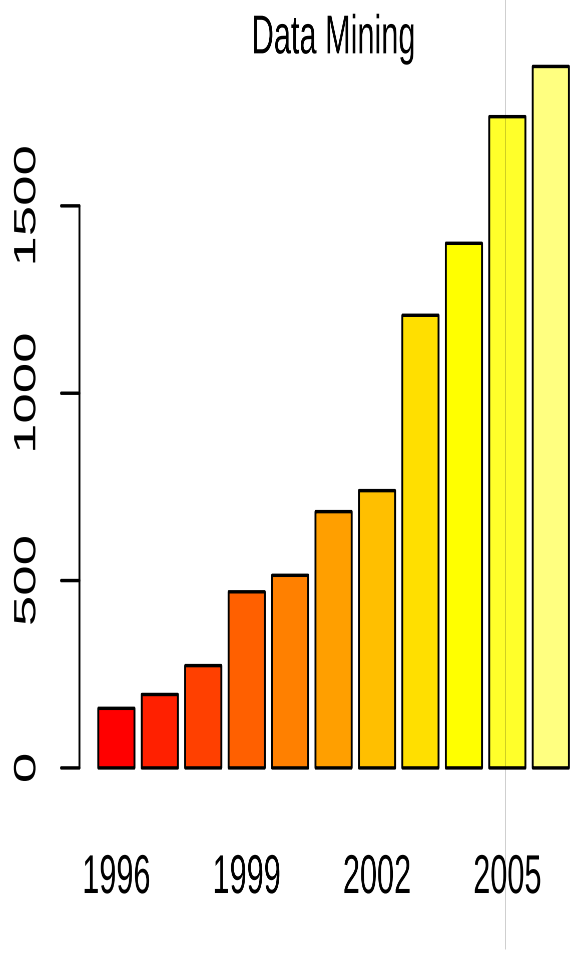
Why Data Mining?

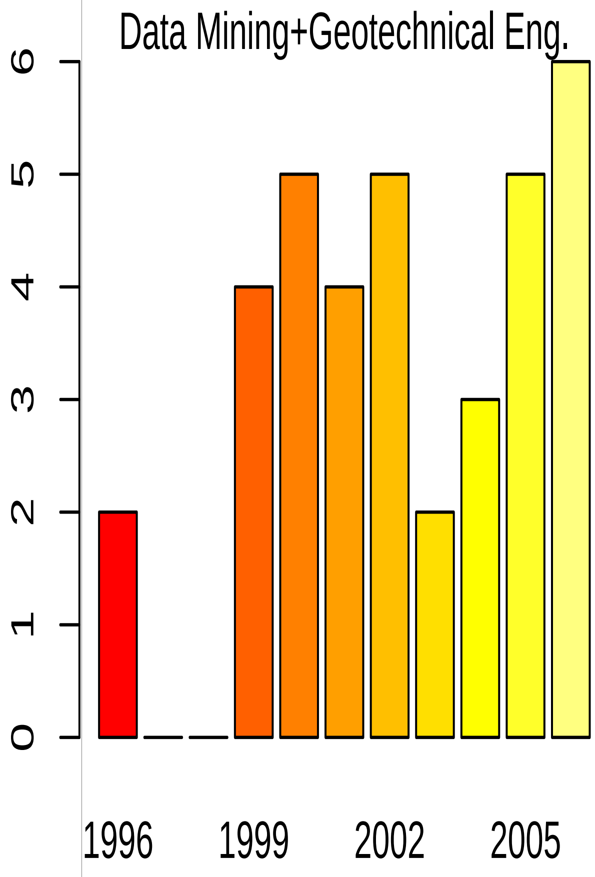
- **Expert Driven Models:** subjective, set by a panel of experts
- **Data Driven Models:** objective (although experts can guide the process), learns from directly data
- In the **Artificial Intelligence** domain, in the 70s there was a great emphasis in **expert systems** (mimic the expert)
- The trend shifted in the 90s to **intelligent systems** (learn from the data or use **hybrid** approaches)

Opportunity: Use Data Mining to create new data driven engineering scores/indexes!

Why Data Mining?

Journal Publications (source ISI Web of Knowledge):





How to do Data Mining?

Software (www.kdnuggets.com):

- Free: WEKA (graphical); R (open source statistical tool);
- Commercial: SAS Enterprise Miner; Clementine (SPSS);

Methodogies:

- SEMMA (SAS)
- CRISP-DM (http://www.crisp-dm.org/):
  - Life cycle with 6 phases: business understanding, data understanding, data preparation, modeling, evaluation, deployment
  - Supported by the industry (SPSS, Daimler-Chrysler, OHRA)

How to do Data Mining?

Data Collection:

- Samples should be **representative**;
- The **more**, the **better** (100, 1000, 10000,…);

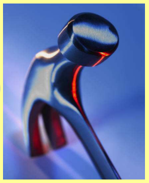
Model Validation (prediction):

- **Holdout:** fit the model with 2/3 of the examples (random sampling), test the model with the rest 1/3;
- More sophisticated validation methods: **10-fold, leave-one-out, bootstrap, …**

Model Description (variable importance, …):

- Apply the description method on all data!

How to do Data Mining?



Selected References:

- U. Fayyad, G. Piatetsky-Shapiro and P. Smyth, Advances in Knowledge Discovery and Data Mining, MIT Press, 1996.
- T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer-Verlag, 2001.
- I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 2005.

Conferences: ACM KDD, IEEE ICDM, ECML/PKDD, ICML, DMin, ICML, …

Journals: SIGKDD Explorations, ACM Transactions on Knowledge Discovery in Data (TKDD), Data Mining and Knowledge Discovery, IEEE Transactions on Knowledge and Data Engineering, Machine Learning, …