Teresa Oliveira and Amílcar Oliveira
{toliveir, aoliveira}@univ-ab.pt

Universidade Aberta and
Center of Statistics and Applications, University of Lisbon

Data Mining and Quality in Service Industry: Review and some applications

2010 IN3 - HAROSA Workshop
November 22-23, 2010 Barcelona, Spain

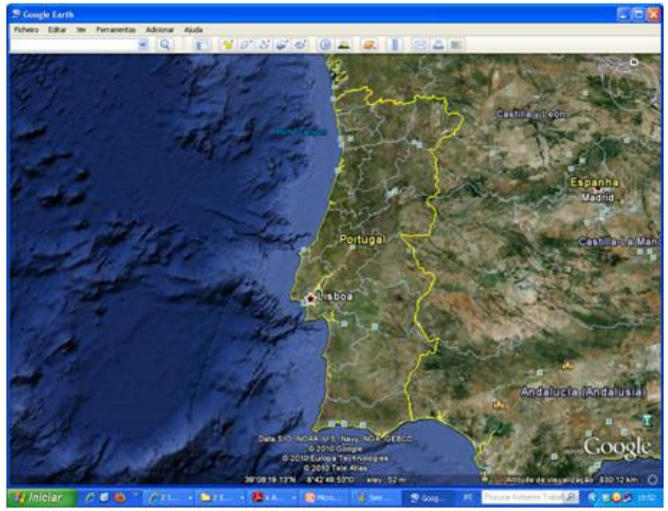
Outline

- Our University
- Our Research Center
- Data Mining : review and applications
- Data Mining and Statistics
- Data Mining and Quality in Service Industry
- Considerations and Future Research
- References

Our University:

UNIVERSIDADE ABERTA

Our University: UNIVERSIDADE ABERTA



- UAb was founded in 1988.
- **Public institution** of higher education dedicated to distance learning in Portugal.
- **10,000 students** in 33 countries over 5 continents.

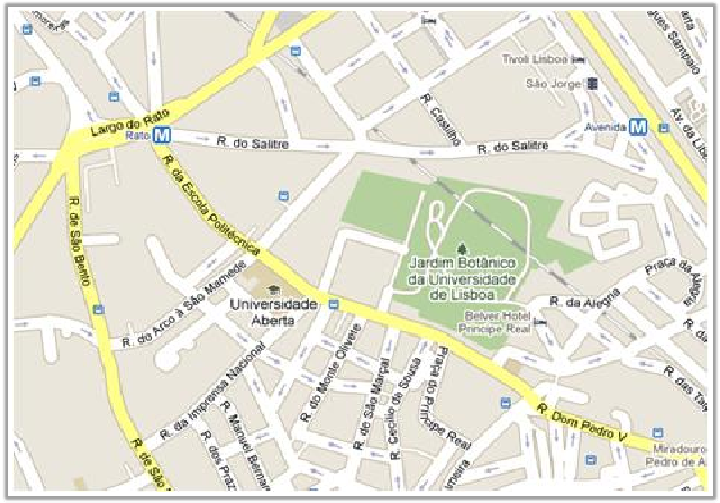
- UAb has already graduated over **9000 students**, among which more than a thousand are master's degrees and about a hundred are doctoral degrees.
- More **830 online courses**.

Our University: UNIVERSIDADE ABERTA

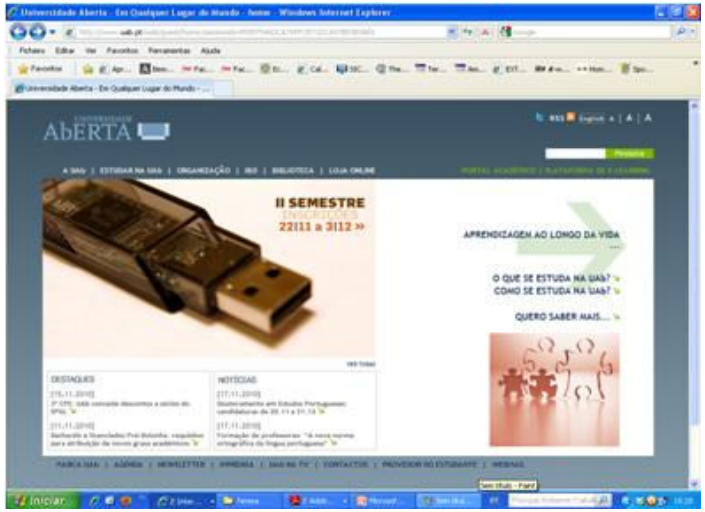
- Pioneer in **e-learning** in Portugal
- Around **200 teachers**, around **300 employees**, headquarters and other service in Lisbon, delegations in Coimbra and in Porto as well as centers in all district capital cities.

Departments

- Education and Distance Learning
- Humanities
- Sciences and Technology
- Social Sciences and Management



Our University: UNIVERSIDADE ABERTA

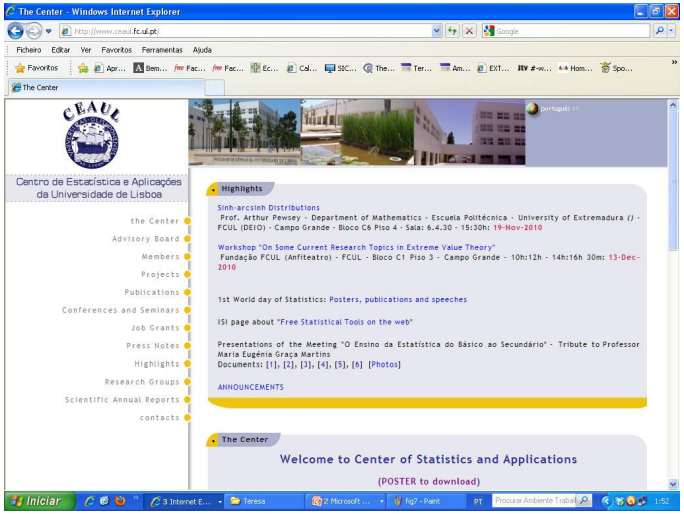


- www.univ-ab.pt
- Virtual pedagogical model
- Learning Management System: **Moodle**

Our Research Center:

Center of Statistics and Applications University of Lisbon

Our Research Center: Center of Statistics and Applications University of Lisbon



- www.ceaul.fc.ul.pt
- CEAUL was established in 1975

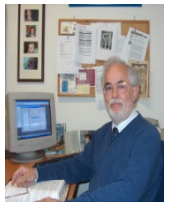
G1– ORDER STATISTICS, EXTREMES and APPLICATIONS

Team Leader: Maria Ivette GOMES

Objectives:

The main objectives of this research team have been in the field of **Statistics of Univariate, Multivariate, Multidimensional and Spatial Extremes**, with special emphasis on their applications to **Life Sciences, Environment, Risk, Insurance, Finance, Experimental Design and Statistical Quality Control.**

Our Research Center: Center of Statistics and Applications University of Lisbon





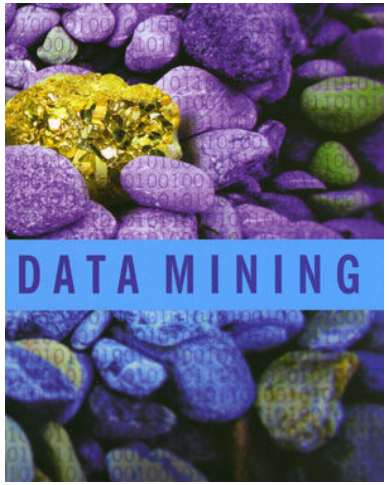
G1 TEAM MEMBERS: Full (Senior; Young); Collaborator
Senior members (Ph.D. prior to 1995)
Young members (Ph.D. pos 2000)

- M. Ivette **Gomes, Fernando Rosado, Luísa Canto e Castro, M. Isabel Fraga-Alves (FC-UL)**
- Laurens de **Haan (TILBURG University -TU)**
- M. Manuela **Neves, Ana Ferreira (ISA-UTL)**
- Fernanda **Figueiredo (FEP-UP)**
- Lígia **Henriques-Rodrigues (ESTT-IPT)**
- M. Cristina **Miranda (ISCA-UA)**
- Teresa Paula **Oliveira , Amilcar Oliveira (UAb)**
- Bjorn **Vandewalle (ISEGI-UNL)**
- Helena **Ferreira, Luísa Amaral, Ana Paula Martins (UBI)**
- M. Graça **Temido (UC)**
- Paulo **Araújo Santos (ESGS-IPS)**
- Paula **Reis (EST-IPS)**
- M. Manuela **Figueira Neves (IPG)**









Data Mining:

Review and Applications

Data Mining: Review and Applications

- Over the last three decades, **increasingly large amounts of data** have been stored electronically and this volume will continue to increase in the future.
- Frawley et al. (1992) present a data mining definition: "*Data Mining is the nontrivial extract of implicit, previously unknown, and potentially useful information from the data*".
- **Data mining is part of the knowledge discovery process** and a current challenge offering a new way to look at data, as an extension of exploratory data analysis.
- Chatfield (1997) argues that the term first appears in a book on econometric analysis dated 1978 and written by E. E. Leamer.
- However there are indications that the term has already been used by statisticians in the mid-sixties of the twentieth century (Kish, 1998).
- Branco, J.A.(2010) mention that the idea behind Data Mining (DM) appears when the **automatic production of data** leads to the accumulation of large volumes of information, which is conclusively established from the mid-decade starting in 1990.

Data Mining: Review and Applications

- It is a process for exploring large data sets in the search for consistent and useful patterns of behaviour in order to lead to the detection of associations between variables and to the distinguishing of new data sets.
- Data mining is then the process of discovering meaningful new correlations, patterns and trends by sifting through vast amounts of data using statistical and mathematical techniques.
- It uses **machine learning**, **neural sets, regression trees and clustering algoritms,** as well **statistical and visualization techniques** to discover and present knowledge in an easily comprehensible form.
- As this knowledge is captured, this can be a key factor to gaining a competitive advantage over competitors in an industry and to improve the quality of the interaction between the organization and their customers.
- To several authors data mining is known as a very important step in the process of Knowledge Discovery in Databases (KDD) which can be represented by the following steps:

Data Mining: Review and Applications

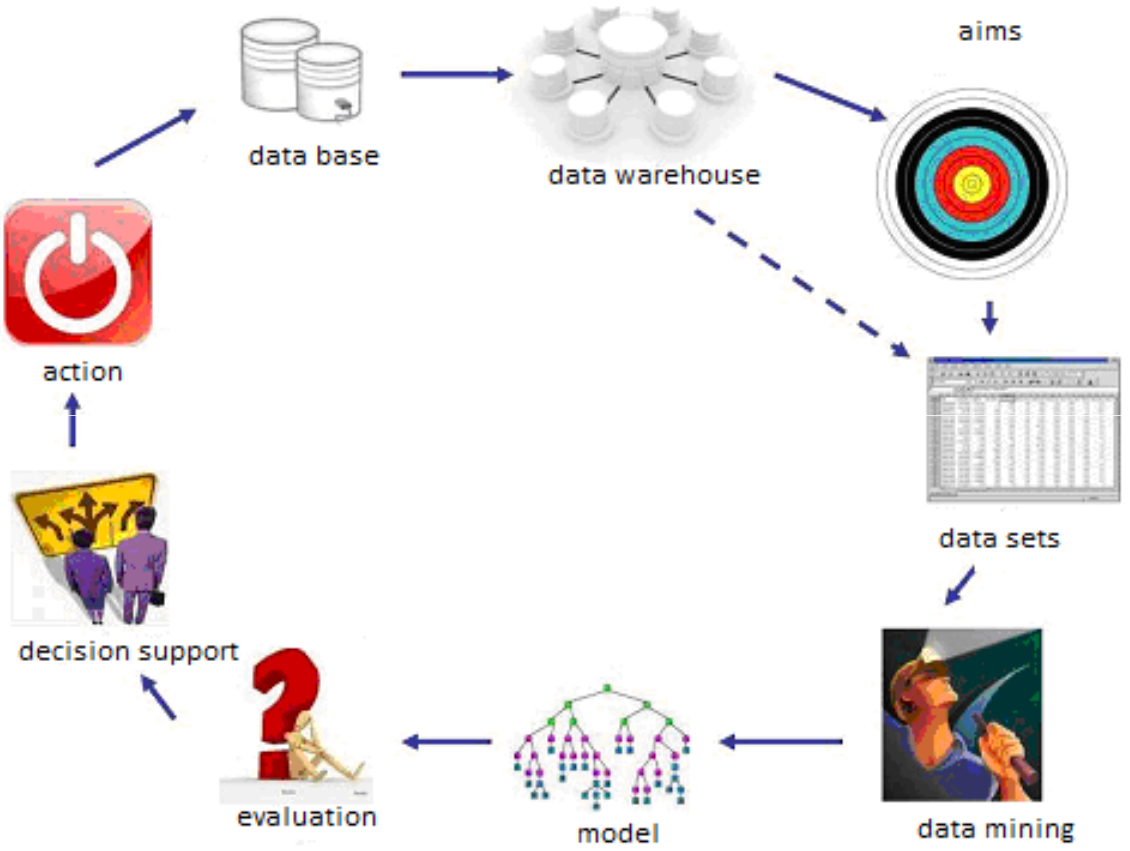
1. Cleaning the data: stage where noise and inconsistent data are removed.
2. Data integration: a stage where different data sources can be combined producing a single data repository.
3. Selection: a stage where user attributes of interest are selected.
4. Data transformation: phase where data is processed into a format suitable for application of mining algorithms (e.g., through operations aggregation).
5. Prospecting: essential step process consisting of the application of intelligent techniques in order to extract patterns of interest.
6. Evaluation or post-processing: stage where interesting patterns are identified according to some criterion of the user.
7. Display of results: stage where techniques are used for representing knowledge in order to present the user the extracted knowledge.

Data Mining: Review and Applications

- Aldana, W.A.(2000) emphasize the difference between DM and KDD. The author refers that KDD concerns itself with knowledge discovery processes applied to databases - it deals with ready data, available in all domains of science and in applied domains such as marketing, planning and control. KDD refers to the overall process of discovering useful knowledge from data while data mining refers to the application of algorithms for extracting patterns and associations from the data.

- Antunes, M.(2010) refer to the website http://www.kdnuggets.com, which centralises information relating to data mining, and where one can see that the consumption sectors, banking and communications represent about 60% of the applications of techniques data mining.

Data Mining: Review and Applications



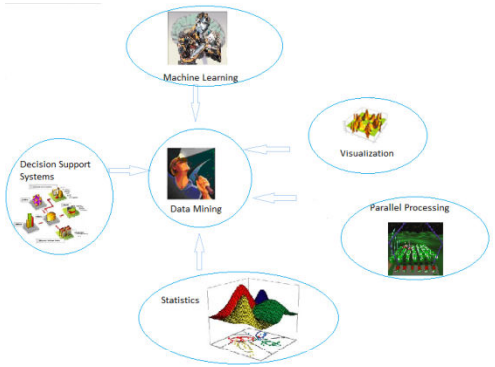
Adapted from Antunes, M. (2010)

Data Mining: Review and Applications

- An overview of the evolution of data mining is shown in the following table:

Evolutionary Step	Enabling Technologies	Characteristics
Data Collection (1960s)	Computers, tapes, disks	Retrospective, static data delivery
Data Acess (1980s)	Relational Data Bases, Structured Query Language	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	Online Analytic processing, multidimensional databases, data warehouses	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging 2000s)	Advanced algorithms, multiprocessor computers, massive databases	Prospective, proactive information delivery

Part from source: An introduction to Data Mining. Pilot Software whitepaper. Pilot Software.1998.



Data Mining and Statistics

Data Mining and Statistics

Statistics play a key rule since Data Mining can be used with three main objectives:

⇒ Explanation: to explain some event or observed measurement, such as why sales of Toshiba laptops are going down;

⇒ Confirmatory: confirming a hypothesis. An insurance company, for example, may want to examine the records of their customers to determine whether two income families are more likely to purchase a health plan than an income families;

⇒ Exploration: to analyze the data looking for new relationships and unexpected. A credit card company can analyze its historical records to determine what factors are associated with people who pose a risk to credit.

Besides many data mining uses many statistical tools, there are some main differences between statistics and Data mining pointed out in the paper by Branco, J.A.(2010):

Data Mining and Statistics

i) Most striking difference is the magnitude and complexity of the data analyzed for DM, unlike the statisticians who usually deals with small data sets static, or clean with small amount of defects and other disorders, constituting a sample in order to answer specific questions concerning the population from which the sample was withdrawal.

ii) Cleaning, compression and data transformation operations are essential initial processing in DM but are not generally required in statistical.

iii) Traditional statistics methods are generally not directly applicable to the treatment of major data sets that DM can analyze.

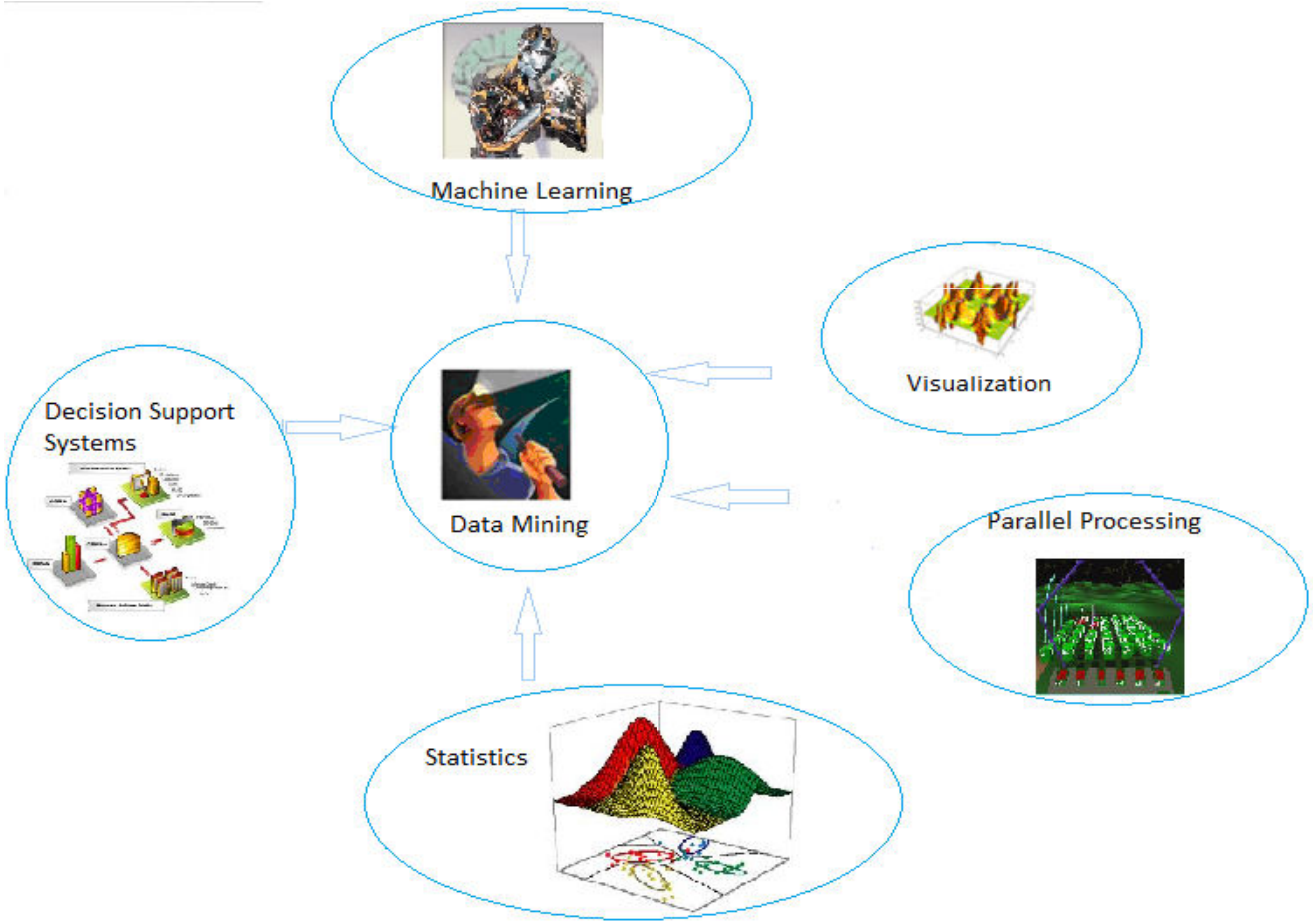
iv) While statistics can be viewed as a process of analyzing relations DM is a process of discovering relationships,

iv) The statistical approach is confirmatory in nature while the DM follows an exploratory approach.

To clarify this distinguishing problem Branco, J.A.(2010) refers to several authors who gave precious contributions, such as Friedman (1998), Hand (1998, 1999a, 1999b), Kuonen (2004) and Zhao and Luan (2006).

Data Mining and Statistics

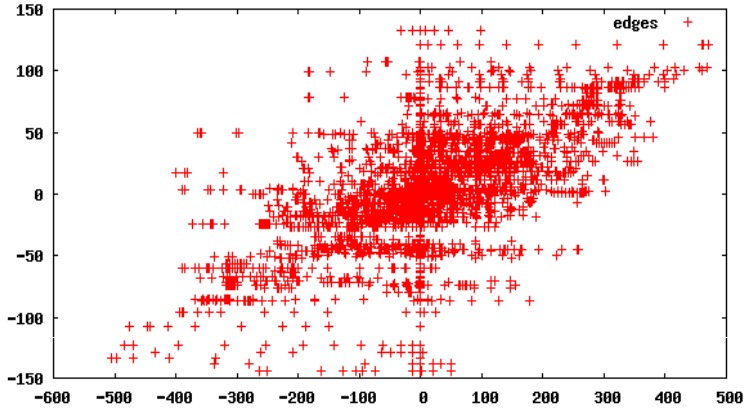
Data Mining uses tools from several major fields:



Data Mining and Statistics

The major statistical techniques used in Data mining are:

Regression - data mining within a regression framework will rely on regression analysis, broadly defined, so that there is no necessary commitment a priori to any particular function of the predictors.

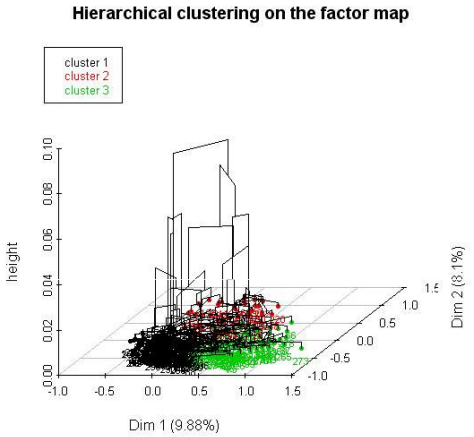


Time Series Analysis - construction of models that detail the behaviour of the data through mathematical methods.

Experimental Design – this technique is a new feature on Data Mining for example in algorithms comparisons.

Data Mining and Statistics

Clustering - In this technique data are grouped according to their classification and group belonging, and these groups are built based on the similarity between the elements of this group.

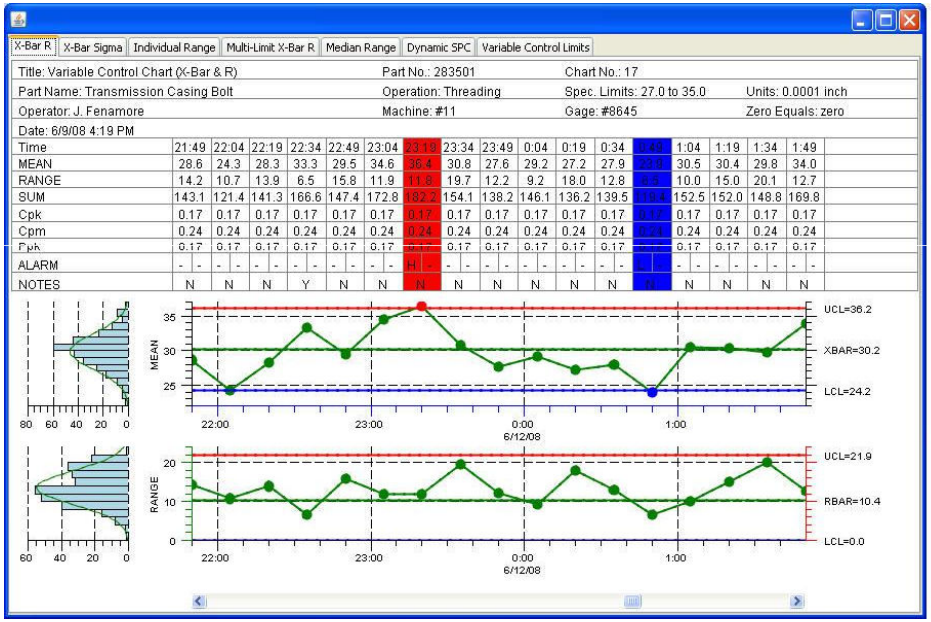


Classification- In data mining this technique is applied to the classification of customers, for example, to classify them according to their social level.

Recognition Patterns - data items are present together with a some probability. For example a supermarket cart: through it one can extract information for the provision of supermarket products to please consumers by placing products close to each other.

Data Mining and Statistics

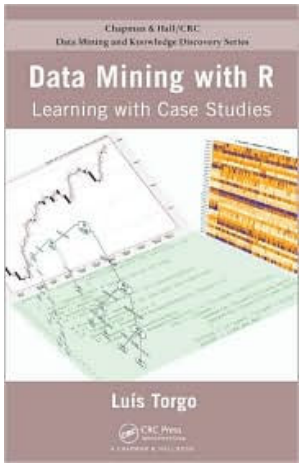
Statistical Quality Control - Modern statistical quality control and improvement include all statistical methods (simple and complex) used to improve manufacturing as well as non-manufacturing processes.



Bayesian Methods - Bayesian approaches are a fundamentally important DM technique. Given the probability distribution, Bayes classifier can probably achieve the optimal result.

Data Mining and Statistics

SOFTWARE ON DATA MINING

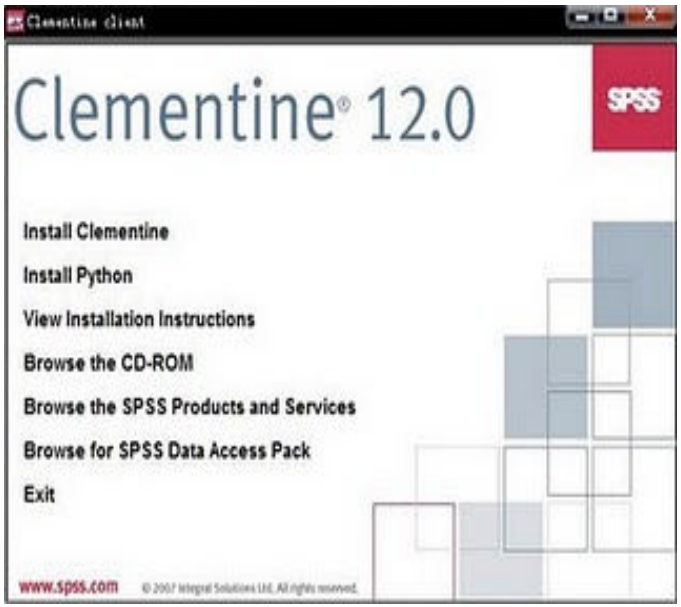


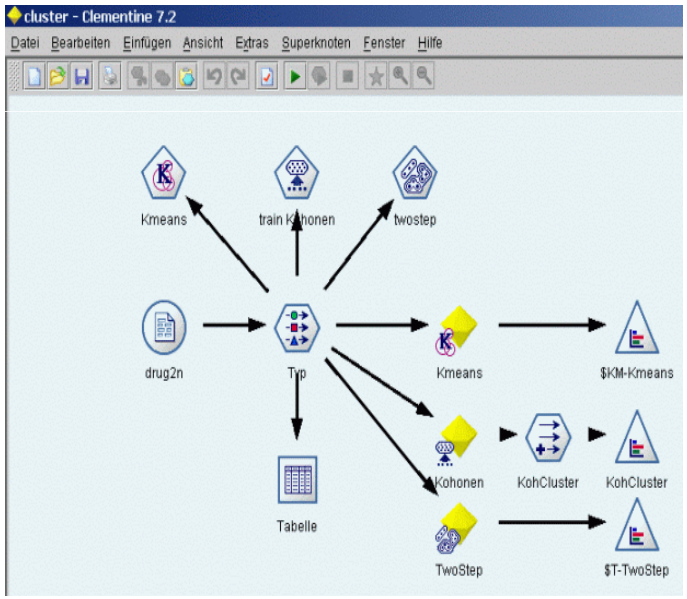
The rule of R GNU project in Statistics is very well known as open source software, freely available, and supported by an international team of specialists in the areas of statistics, computational statistics and computer science. Now-a-days R is also increasingly being recognised as providing a powerful platform for data mining. The Rattle package, http://rattle.togaware.com in R is dedicated to data mining and it is widely used by in industry, by consultants and for the teaching world wide.

Data Mining and Statistics

SPSS – Package Clementine

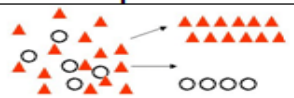
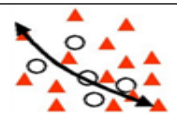
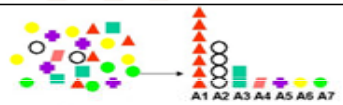
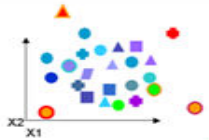
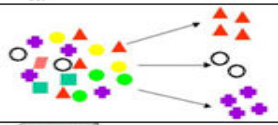
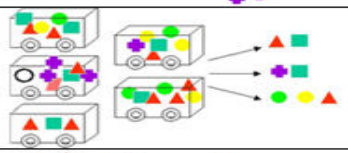
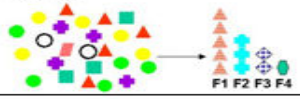
http://credit-new.com/software/1681-spss-clementine-v12.html





Data Mining and Statistics

Oracle Data Mining (Oracle Corporation 2007)

Technique		Algoritm
Classification		• Logistic Regression • Naive Bayes • Support Vector Machine • Decision Tree
Regression		• Multiple Regression • Support Vector Machine
Attribute importance		• Minimum Description Length
Anomaly Detection		• One-Class Support Vector Machine
Clustering		• Enhanced K-Means • Orthogonal Partitioning Clustering
Association Rules		• Apriori
Feature Extraction		• Non-negative Matrix Factorization

Models in Oracle

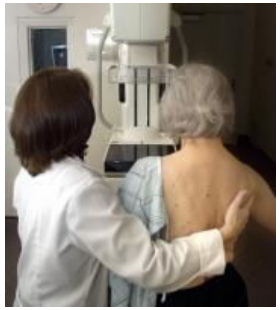
Data Mining and Statistics

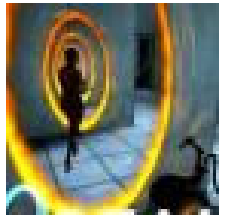
Data Mining with Statistica

Data Mining and Quality in Service Industry

Data Mining and Quality in Service Industry

- **Data Mining is ready for application in industry services** and in the business community since it is supported by the following mature technologies: Massive data collection, powerful multiprocessor computers and data mining algorithms, as we can see in Aldana, W.A.(2000) .





http://topnews.co.uk/

Data Mining and Quality in Service Industry

Service Industry and Data Quality

- The purpose of **service industry collections is to provide information on the size, structure and nature** of the industry under study. Data collections can have origin in several areas, for example the information on business income and expenses and the employment status.

- Data mining is used for a variety of purposes in **private and public services** as pointed by Seifert, J.W.(2004). The author mentions that industries such as banking, insurance, pharmacy, medicine and retailing, among others, commonly use this technique to reduce costs, enhance research and increase sales.

Examples:

- **Banking and insurance**: to detect fraud and assist in risk assessment, like credit scoring
- **Pharmacy**: using data on chemical compounds and genetic material to help guide research on new treatments
- **Medical community**: to predict the effectiveness of a procedure or medicine
- **Retailers**: to analyse the effectiveness of product selection, coupon offers, products purchased together.
- **Telephone companies**: to assess which customers are likely to remain subscribers.

Data Mining and Quality in Service Industry

Service Industry and Data Quality

- In the paper by Farzi and Dastjerdi (2010) the authors introduce a **method for measuring the data quality based on data mining algorithms.**

This algorithm has three steps, which calculates the data quality of transaction.

Step1: Extract association rules, which depend on input transaction and are adapted by the functional dependency.

Step2: Separate compatible and incompatible association rules.

Step3: Calculate the quality of input transaction.

Data Mining and Quality in Service Industry

Service Industry and Data Quality

- Ograjensek (2002) emphasizes that to provide a satisfactory definition of services, some authors (e.g., Mudie and Cottam 1993; Hope and Mühlemann,1997; Kasper et al., 1999) characterise them with these important features:

- ***Intangibility***: services cannot generally be seen, tasted, felt, heard or smelled before they are bought.

- ***Inseparability***: services are produced and consumed at the same time.

- ***Variability***: the quality of the same service may vary depending on who provides it as well as when and how it is provided.

- ***Perishability***: services cannot be stored for later sales or use; lack of demand cannot be evened out by producing to an inventory.

Data Mining and Quality in Service Industry

Quality in Service Industry

- **Quality of a service is the degree to which the bundle of services attributes as a whole satisfies the customer**, Murdick et al. (1990); it involves a comparison between customer expectations (needs, wishes) and the reality (what is really offered).

- According to Juran (1999), quality can be defined as "fitness for use".

- Whenever a person pays for a service or product the main aim is in getting quality.

There are two main reasons why customers must be given quality service:

⇒ Industry is very competitive and customers have a huge variety of alternatives;
⇒ most customers do not complain about meeting problems, they simply go and make business elsewhere.

- In a quality service the special challenge is the need to meet the customer requests while remaining the process economically competitive.

Data Mining and Quality in Service Industry

Quality in Service Industry

- To **ensure quality assurance of the companies, there are the shadow shopping companies**, who act as the companies quality control laboratories, developing experiments, drawing inference and providing the solutions for improvement. In the case of the service industry the quality assurance service is provided by the various surveys and inspections that are carried out from time to time.

- **Modern statistical quality control and improvement** include all statistical methods used to improve manufacturing and non-manufacturing processes. In order to improve quality in the service sector it is important to realise that every process generates information that can be used for its improvement.

- **Designing quality into products and processes is a top priority** and, sophisticated statistical methods such as **statistical experimental designs** may be more appropriate for attaining better (optimum) values of important quality characteristics of products and processes (Box and Bisgaard, 1987).

Data Mining and Quality in Service Industry

Quality in Service Industry

- Ograjensek (2002) presents, among others, the following example to show how experimental methods can also be used in the service sector.

The author mention that factors influencing service quality can be deliberately changed. The effects of changes have to be closely examined to determine how they affected the service or the service delivery process.

- **Example**: On-line retailers are beginning to experiment with the design of their commercial web sites, trying to increase the number of visits and –consequently – the number of sales. There are examples of companies using their on-line catalogue to test price elasticity of its customers.

Data Mining and Quality in Service Industry

Quality in Service Industry

- Ograjensek (2002) refers to **statistical quality control and to Six-Sigma initiative**, as it is "a program aimed at the near-elimination of defects from every product, process and transaction" (Hoerl, 1998). The author mention that it is a disciplined quantitative approach for improvement of defined indicators (called "metrics") in all types of business processes.

- **Six-Sigma was initially introduced by Motorola** and widely used in giants such as General Electric and the initiative has a broad business character: quality improvements are seen as ground stones for huge cost savings .

- Sigma (σ) is used by statisticians to measure the variability in any processes. A company's performance is measured by sigma level of their business processes. Traditionally companies accept three or four sigma performance levels as a norm.

- **The Six Sigma is a response to the increasing expectations of customers and the increased complexity of modern products and processes, Trnka, A. (2010).**

Considerations and Future Research

- A key challenge for data mining companies in the current century is **the creation of real-time algorithms for web mining analysis** in this Internet age.

- The classical statistical were developed and prepared to study the small ensembles and it seems to be frequently inappropriate concerning large sets analysis.

- It urges in future modern statistics the creation of appropriate methods, and to continue their development to meet the challenges and new issues that have put large sets.

- It is a fact that **data mining is not complete without statistics and statistical techniques need the Data Mining** on improving the analysis of large data sets.

- Since Statistics and Data Mining, although distinct, share objectives and methods, our conviction is that the cooperation between specialists from both areas provides an advance more rigorous, more secure and faster.

Considerations and Future Research

- The use of more sophisticated statistical methods can be facilitated by modern user-friendly menu driven statistical software.

- There is the need for professionals with skills required to use experiments successfully. Without knowing the subject matter, the experiment or survey planning and data interpretation are of limited use and very likely will not have the desired impact.

- Academia should play a vital role in these endeavours, particularly in specialist training with these skills.

References

- Aldana, W.A. (2000). Data Mining Industry: Emerging Trends and New Opportunities. Master Thesis, MIT.
- Antunes, M.(2010). CRM e Prospecção de Dados – ao seu service. Boletim da Sociedade Portuguesa de Estatística, Primavera de 2010,. Editor Fernando Rodado. Pag.34-39.
- Box, G.E.P. and Bisgaard, S. (1987): The Scientific Context of Quality Improvement. A Look at the Use of Scientific Method in Quality Improvement. *Quality Progress*, June, 54-61.
- Branco, J.A.. Estatísticos e mineiros (de dados) inseparáveis de costas voltadas? Boletim da Sociedade Portuguesa de Estatística, Primavera de 2010,. Editor Fernando Rodado. Pag.40-43.
- Berry, M.J.A., Linoff, G.S. *Data Mining Techniques For Marketing, Sales, and Customer Relationship Management*. 2004. Wiley Publishing, Inc.
- Chatfield, C. (1997). Data mining. *Royal Statistical Society News,* 25, 3, 1-2.
- Christmann, A. e Shen, X. (2009). Editorial: On the interface of statistics and machine learning. *Statistics and Its Interface*, 2, 255-256.
- Farsi, S. and Dastjerdi, A.B. (2010). Data Quality Measurement Using Data Mining. International Journal of Computer Theory and Engeneering, Vol.2,nº1, 115-118.
- Frawley, W., Piatetsky-Shapiro. Mathews, C. (1992). Knowledge discovery in Databases: an overview. *AI Magazine*, 213-228.
- Friedman, J. H. (1998). Data mining and statistics: what is the connection. Unpublished. www-stat.stanford.edu/~jhf/ftp/dm-stat.ps.
- Ganesh, S. (2002). Data mining: should it be included in the statistics curriculum? *The 16 th International Conference on Teaching Statistics (ICOTS 6).* Cape Town, South Africa.
- Goodman, A. (2001). Statistics is the road from data mining to knowledge discovery. *KDnuggets*: *News*: n24: item1.
- Hand, D. J. (1998). Data mining: statistics and more? *The American Statistician*, 52, 112-118.
- Hand, D. J. (1999a). Statistics and Data mining: intersecting disciplines. *SIGKDD Explorations*, 1, 16- 19.
- Hand, D. J. (1999b). Data mining: new challenges for statisticians. *Social Science Computer Review* 18, 442-449.
- Hand, David J.; Mannila, Heikki; Smyth, Padhraic. *Principles of Data Mining*. 2001. MIT Press.
- Hope, C. and Mühlemann, A. (1997): *Service Operations Management.Strategy, Design and Delivery*. London: Prentice Hall.
- Hoerl, R.W. (1998): Six Sigma and the Future of the Quality Profession. *Quality Progress*, June Ed..
- Juran, J.M. (1999): How to think about Quality. In J.M. Juran, A.B. Godfrey, R.E. Hoogstoel, and E.G., Schilling (Eds.): *Quality-Control Handbook*. New York: McGraw-Hill
- Kasper, H., van Helsdingen, P., and de Vries, W. Jr. (1999): *Services Marketing Management. An International Perspective*. Chichester: John Wiley & Sons.
- Kish, L. (1998). Royal Statistical Society News, 25, 6, 8.
- Kuonen, D. (2004). Data mining and statistics: what is the connection? *The Data Administrative Newsletter*. Switzerland.
- Mudie and Cottam (1993) :*The Management and Marketing of Services*. Oxford: Butterworth-Heinemann.
- Murdick, R. G., Render, R., Russell, R. S., (1990), *service operations management*. Prentice-Hall, Inc.
- Ograjensek (2002). Applying Statistical Tools to improve quality in the service sector. Metodoloski zvezki, 18. Ljubljana:FDV.
- Seifert, J. W. (2004). "Data Mining: An Overview." CRS Report RL31798.
- Torgo, L.C.(2010). Data Mining with R: Learning with case studies. Chapman & Hall.
- Trnka, A. (2010).Classification and Regression Trees as a Part of Data Mining in Six Sigma Methodology.

Proceedings of the World Congress on Engineering and Computer Science 2010 Vol I

WCECS 2010, October 20-22, 2010, San Francisco, USA

- Zhao, C. M. e Luan, J. L. (2006). Data mining: Going beyond traditional statistics. *New Directions for Institutional Research*, 131, 7-16