Knowledge Discovery from Data Bases
Proposal for a MAP-I UC

P. Brazdil[1], João Gama[1], P. Azevedo[2]
[1]Universidade do Porto; [2]Universidade do Minho;

1 Knowledge Discovery from Data Bases

"We are deluged by data: scientific data, medical data, demographic data, financial data, and marketing data. People have no time to look at this data. Human attention has become the precious resource. So, we must find ways to automatically analyze the data, characterize trends in it, and automatically flag anomalies." (Han e Kamber, 2006).

The development of information and communication technologies make possible collect data with high degree of detail, that might be automatically transmitted at high-speed. Some examples of real-world applications include: TCP/IP traffic, queries in search engines over Internet, records of telecommunication calls, SMS, emails, stock market, sensors in electrical grid, etc. For illustrative purposes, we present some numbers: The number of daily phone calls is around 3 billions, the number of SMS is 1 billion daily, the number of sent emails is around 30 billions.

Most of this information will never be seen by a human being. Taking this into account, tools for automatic real-time data analysis are of increasing importance. The computer process, analyze, and filter the data, selecting the most promising hypothesis. Some typical applications include: user modeling, activity monitoring, sensor networks, classification, intrusion detection, etc.

2 Team

- Pavel Brazdil, Professor Catedrático, Faculdade de Economia, Universidade do Porto;
  Prof. Dr. Pavel Brazdil got his PhD degree from the University of Edinburgh in 1981. The thesis was in he area of Machine Learning.

In late 70's, when this work was carried out, there was relatively little work done in this area. In 1996 he obtained habilitation at the University of Porto and since 1998 is Full Professor. Currently he is the Coordinator of R&D Unit LIAAD, which was earlier known as group NIAAD of LIACC (he was one of its founders in 1988). Pavel Brazdil is known for his activities in Machine Learning, Data Mining, Meta-learning and Text Mining. He has participated in two international projects and was a technical coordinator of one of them (METAL, 5th FP) and participated in various international research networks. He has supervised 9 PhD students all of whom have completed their studies. He has organized more than 10 international conferences or workshops and participated in many program committees. He is a co-author / co-editor of 5 international books and has published more than 130 articles. Since 2007 he is a Fellow of ECCAI (European Coordinating Committee for Artificial Intelligence.

- João Gama, Professor Associado, Faculdade de Economia, Universidade do Porto;
  João Gama is a researcher at LIAAD, the Laboratory of Artificial Intelligence and Decision Support of the University of Porto, working at the Machine Learning group. His main research interest is in Learning from Data Streams. He has published several articles in change detection, learning decision trees from data streams, hierarchical Clustering from streams, etc. Editor of special issues on Data Streams in *Intelligent Data Analysis*, *J. Universal Computer Science*, and *New Generation Computing* Co-chair of a series of Workshops on Knowledge Discovery in Data Streams, ECML 2004, Pisa, Italy, ECML 2005, Porto, Portugal, ICML 2006, Pittsburg, US, ECML 2006 Berlin, Germany, SAC2007, Korea, and the ACM Workshop on Knowledge Discovery from Sensor Data to be held in conjunction with ACM SIGKDD 2007. He edited the books *Learning from Data Streams-Processing Techniques in Sensor Networks*, published by Springer, and *Knowledge Discovery from Sensor Data*, published by CRC. He served as program chair at ECML 2005, ADMA 2009, and Conference chair at DS 2009.

- Paulo Azevedo, Professor Auxiliar, Escola de Engenharia, Universidade do Minho
  Paulo Jorge Azevedo holds a PhD in Computer Science (Imperial College, University of London - 1995) and a MSc in Information Technology (Imperial College, 1991). He is an Assistant Professor at the Department of Informatics of the University of Minho, where he tea-

ches informatics to undergraduates and data mining and data analysis related courses to post-graduate students. His research is concentrated in the fields of Machine Learning, Data Mining and its applications to Bioinformatics problems. He was the coordinator of the national FCT funded project CLASS and he currently participates in the also FCT funded projects Site-O-Matic, on web automation and P-found and ProtUnf on the analysis of induced Molecular Dynamic Simulations of Protein Unfolding. He is currently supervising several PhD and MSc students in the areas of Data Mining and Bioinformatics. Paulo Azevedo was member of several program committees, among others the PKDD 2005 conference, Principles and Practice of Knowledge Discovery and Data Mining, and the EPIA-01, 03, 05 and 09 (the Portuguese Conference on Artificial Intelligence), DS'09 (International Conference on Discovery Science) and ECML'09 (European Conference on Machine Learning). He co-organized the Workshop on Computational Methods in Bioinformatics under EPIA'2005. He has also been vice-chair of the Portuguese Society for Artificial Intelligence from 2000 to 2003

3 Main Goals

At the end of the semester the students should be able to:

1. Design of multidimensional data bases;
2. Identify the basic tasks in knowledge extraction from data bases;
3. Identify and use the main methods and algorithms for knowledge representation;
4. Apply the main methods and algorithms for each mining task;
5. Apply the main methods and algorithms in real-world problems and adapt to new contexts.

4 Program

- Knowledge Discovery in Data Bases
- Data warehouses and OLAP
  - Schemes for multidimensional data bases;
  - From OLAP to *On-Line Analytical Mining*

- Cluster Analysis
  - Cluster Analysis: concepts and methods;
  - Partitioning Methods;
  - Hierarchical Methods;
  - Incremental Methods;
- Association Analysis
  - Frequent pattern mining;
  - Algorithms;
  - Pos-processing;
  - Applications;
- Predictive Data Mining: Classification and Regression.
  - Optimization Methods;
  - Probabilistic Methods;
  - Search based Methods;
- Evaluation in Predictive Data Mining.
  - Evaluation: goals and perspectives;
  - Loss Functions;
  - Sampling Methods;
  - Bias-Variance analysis;
  - Cost-benefit analysis;
- Pré-processing
  - Data summarization;
  - Data cleaning;
  - Feature selection;
- Ensembles and Multiple Models
  - Concepts and methods;
  - Combining Homogeneous Models;
  - Combining Heterogeneous models;

5 Teaching Methods and Evaluation

The teaching method consists of theoretical-practical classes. The evaluation consists of home-works and a final exam.

6 Bibliography

Recommended books:

- Data Mining, Concepts and Techniques, Jiawei Han e Micheline Kamber, Morgan Kaufmann, 2006
- Principles of Data Mining, D. Hand, H. Mannila, P. Smyth; The MIT Press, 2002
- Machine Learning, Tom Mitchell; McGraw Hill, 1997.
- Intelligent Data Analysis, Michael Berthold e David Hand; Springer, 1999.

Other books of interest:

- *Learning from Data Streams - Processing Techniques in Sensor Networks*, J. Gama, M. Gaber; Springer; 2007.
- *Introduction to Data Mining*; Pang-Ning Tan, Michael Steinbach e Vipin Kumar; Addison-Wesley; 2006.
- *Pattern Recognition and Neural Networks*, Ripley, B.D.; Cambridge University Press, 1996.
- *Computer Systems that Learn*, S. Weiss, C. Kulikowski; Morgan Kaufmann, 1994.
- *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, I. Witten and E. Frank; Morgan Kaufmann, 2000.
- *Sistemas de Suporte a Decisão*, Bruno Cortes, FCA-Editora de Informatica, 2005
- *Data Mining, Descoberta de Conhecimento em Bases de Dados*, M. Filipe Santos, Carla Azevedo, FCA-Editora de Informatica, 2006

7 Software

The use of software tools has the main goal of solving practical problems. The study, analysis, and evaluation in small-scale applied problems as a formative perspective. We choose two software tools, frequently used in data mining teaching:

- R (Ihaka e Gentleman, 1996)
  R is a statistical oriented programming language. The interface is command line.

- WEKA (Witten e Frank, 2005)
  Weka is a machine-learning oriented software. It uses a graphical interface, with the possibility to develop sequences of tasks. The Knowledge Explorer permit to decompose a complex problem into sub-problems in a graphical environment.

Referências

Han, J. e Kamber, M. (2006). *Data Mining Concepts and Techniques*. Morgan Kaufmann.

Ihaka, R. e Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314.

Witten, I. H. e Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.