Knowledge Discovery from Data Bases
Proposal for a MAPi UCT
Unidade Curricular em Tecnologias

João Gama   Alipio Jorge   Rita P. Ribeiro
jgama@fep.up.pt   amjorge@fc.up.pt   rpr@fc.up.pt

Paulo Azevedo
pja@di.uminho.pt

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 and 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 billion; the number of SMS is 1 billion daily, the number of sent emails is around 30 billion.

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 processes, analyze, and filter the data, selecting the most promising hypothesis. Some typical applications include: user modeling, activity monitoring, sensor networks, classification, intrusion detection, etc.

Scientific areas: Data Mining, Machine Learning, Computer Science.

1.1 Main Goals

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

1. Formulate a decision problem as a data mining problem;

2. Identify the basic tasks in knowledge discovery from data bases;

3. Identify and use the main methods in solving data mining problems;

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.

1.2 Team

- Alipio Jorge, U. Porto

- João Gama, U.Porto (coordinator)

- Paulo Azevedo, U. Minho

- Rita P. Ribeiro, U. Porto

1.3 Syllabus

- Introductory Concepts

  - Data Mining tasks;
  - Unsupervised Learning (Hand et al., 2001)
    * Cluster Analysis
    * Association Analysis
  - Predictive Supervised Learning Bishop (1995)
    * Classification
    * Regression
    * Deep Learning
  - Evaluation in Predictive Data Mining (Japkowicz and Shah, 2011).
  - Ensembles and Multiple Models (Gama et al., 2017)

- Advanced Topics

  - Meta-learning, AutoML

- Social Network Analysis
    * Concepts and methods;
    * Evolution of Networks;
- Text Mining and Natural Language Processing
    * Concepts and methods;
    * Information retrieval;
    * Document classification;
- Web Mining and Link Analysis (Han et al., 2012)
    * Concepts and methods;
    * Web and Structure mining;
    * Link analysis;
- Big Data and Data stream Mining (Gama, 2010)
    * Big Data: Applications and tools
    * Concepts and methods;
    * Summarizing data streams;
    * Knowledge discovery from data streams;
- Advanced Topics in classification (Branco et al., 2016)
    * outliers, rare cases, novelty detection,
    * structured output prediction.
- Data Mining Standards and Processes

1.4 Teaching Methods and Evaluation

The teaching method consists of theoretical-practical classes. The evaluation consists of home-works.

1.5 Bibliography

Recommended books:

- *Extração de Conhecimento de Dados - Data Mining*, J. Gama, A. Carvalho, K. Faceli, A. Lorena, M. Oliveira; Sílabo, 2012.

- Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, Jean Pei; Morgan Kaufmann, 2014.

- *Knowledge Discovery from Data Streams*, J. Gama, CRC Press, 2010

- *Principles of Data Mining*, D. Hand, H. Mannila, P. Smyth; The MIT Press, 2002

- *Guide to Intelligent Data Analysis: How to Intelligently Make Sense of Real Data*, Michael R. Berthold, Frank Klawonn, Frank Hoppner, Christian Borgelt, Springer, 2010

1.6   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:

- Python (Van Rossum and Drake, 2009) the most common used language for Data Mining

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

- Knime (Adä and Berthold, 2010) or Rapid-Miner - new generation of data mining tools.

- WEKA (Witten and Frank, 2005) is s machine-learning oriented software. Large set of machine learning algorithms available.

Referências

Adä, I. and Berthold, M. R. (2010). The new iris data: modular data generators. In *KDD*, pages 413–422.

Bishop, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Branco, P., Torgo, L., and Ribeiro, R. P. (2016). A survey of predictive modeling on imbalanced domains. *ACM Comput. Surv.*, 49(2):31:1–31:50.

Gama, J. (2010). *Knowledge Discovery from Data Streams*. Chapman and Hall / CRC Data Mining and Knowledge Discovery Series. CRC Press.

Gama, J., Carvalho, A., Faceli, K., and Lorena, A. (2017). *Extração de Conhecimento de Dados - Data Mining*. Sílabo, 3 edition.

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

Han, J., Kamber, M., and Pei, J. (2012). *Data mining concepts and techniques, third edition.* Morgan Kaufmann Publishers.

Hand, D., Mannila, H., and Smyth, P. (2001). *Principles of Data Mining.* Bradford Books, MIT Press, MA.

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

Japkowicz, N. and Shah, M., editors (2011). *Evaluating Learning Algorithms: A Classification Perspective.* Cambridge University Press.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual.* CreateSpace, Scotts Valley, CA.

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