A Data Mining E-learning Tool: Description and Case Study

Francesco Maiorana, Angelo Mongioj, and Marco Vaccalluzzo

***Abstract*—Data analysis and data mining techniques are recognized as important disciplines that deliver 'sophisticated domain knowledge'. The Association of Computing Machinery (ACM) and the IEEE-Computer Society suggest in the 2013 draft Computer Science curricula, Data Mining as an elective course and an important course topic in previous Information Systems or Information Technology curricula. In this paper we report a four year teaching experience in an Information Systems course for students in Management Engineering in which basic Data Mining techniques were taught as part of the course. The experience led to the design of an e-learning web platform, based on Matlab which is useful for data mining studies. A case study is presented in the field of customer switching prediction. On the basis of customer usage and demographic data the analysis aims to build a model to identify which customers are likely to switch from a 2G to a 3G network. The analysis was performed using a publicly available dataset.**

***Index Terms*— E-learning, Data mining, Matlab**

I. INTRODUCTION

DEVELOPMENT of data mining algorithms and tools has become an important research topic. Several tools, both commercial and open source, are available and they can help in the task at hand. For a recent review of the available tools the reader can refer to [1].

Recently, parallel data mining algorithms have been developed on a grid [2] and in cloud architecture [3] with applications in many domains such as bioinformatics. For example in [4] the authors present the application of an unsupervised clustering algorithm to classify documents in a vector space model representation and automatically extract novel associations. The influence of the input noise on the classification accuracy in a grid architecture has been studied in [5]. Visualization tools such as the one described in [6] are also of major importance.

The importance of data mining algorithms and techniques has been also recognized in modern curricula. The Association of Computing Machinery (ACM) and the IEEE-Computer Society suggest in the 2013 draft computer science curricula [7] that 'data mining' should be an elective course in the Information Management curriculum and an important topic in many courses in the Intelligent Systems curricula. The data mining topics are widely recognized also in other curricula such as Computer Engineering, Information Technology and so on.

The field encompasses a broad spectrum of knowledge ranging from machine learning, statistics, artificial intelligence and so on. Moreover, the subject is taught in different classes with a different background of knowledge.

For these reasons a tool that facilitates the data mining process in order to focalize on algorithm selection and results analysis can be helpful, especially for students without a strong background in algorithms and programming.

In this paper we report a four year teaching experience on a Master course in Information Systems where we taught data mining techniques for half of a 60 hour course. All the enrolled students had a background in management engineering with no or little experience in programming. As data mining tool we chose Matlab, after a tutorial on its use and programming.

At the conclusion of the teaching experience we designed a web tool for e-learning that, making leverage on a web interface, can assist students in data analysis and model construction by offering a web interface that supports the application of the basic data mining algorithms and assists in model construction. Also, the web interface allows the students to focus on parameter tuning and analysis of the results.

To illustrate how the methodology of teaching data mining works in practice, in the paper we report a case study in the domain of prediction of customer switching behavior [8] that has been carried out on the basis of a publicly available dataset, used in the PAKD data mining competition [9], which collects data of a telephone company. In particular, the students were asked to construct a model to predict, on the basis of customer usage and demographic data, which customers are likely to switch from a 2G to a 3G network.

This paper is organized as follows: section 2 briefly describes the teaching experience and presents the software architecture of the e-learning tool for data mining, section 3 describes the case study of the PAKDD data analysis, and section 4 draws the conclusions and highlights future work.

II. A DATA MINING E-LEARNING PLATFORM

The e-learning platform was designed after a four year teaching experience of data mining concepts in an Information Systems Master course in Management

Manuscript received March, 18, 2012, revised April, 12, 2012..

Francesco Maiorana is with Department of Electrical, Electronic and Computer Engineering, University of Catania, Viale Andrea Doria, 6 – 95125 Catania, Italy, phone: 39-95-7382372; fax: 39-95-7382397; e-mail: Francesco.maiorana@dieei.unict.it.

Angelo Mongioj is with ENI - Refining and Marketing Via Laurentina, 449 - 00142 Roma, Italy.

Marco Vaccalluzzo is with Ernst & Young Financial-Business Advisors Italia, Via Wittigens, 6 – 20123 Milano, Italy.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Engineering. The students in the course came from different backgrounds with no, or minimal, programming experience. None of the students had prior knowledge of Matlab. The majority of the students in the same semester attended a course in statistics where Matlab and the statistical toolbox were used. Almost half of the information systems course dealt with data mining techniques.

In particular, after a Matlab tutorial, the subjects taught, according to [10] and [11], were: exploratory data analysis, uni, bi and multivariate analysis, distance and similarity measures, hierarchical classification, decision trees, neural networks, support vector machines, nearest neighbor models and finally methods to evaluate data mining results.

The students were requested to present a final project at the end of the case study where they had to perform a complete data mining workflow from data analysis and cleaning to model construction, testing and evaluation using the Matlab software suite and its toolbox.

As to leverage programming details, especially in the initial learning path, at the conclusion of the four year teaching experience, we explored the possibility of designing an in house e-learning web based environment by using the Matlab web server.

The aims of the e-learning platform were:

- Allowing the students to register and to enter into the archive of the most relevant case studies,
- Leveraging the use of the Matbal code by setting the principal parameters of the most important functions of the model construction phase via a user-friendly web interface for a first, rapid data exploration and parameter fine tuning
- Tracing student progress using the Matlab tools, which are organized into thematic areas resembling the course architecture
- Offering help and feedback

The software architecture is shown in figure 1 as well as the initial interface shown on the left.

The software architecture is composed of the following main parts: a web browser client, the Apache web server, the Matlab server (responsible for the communication between the Web application and Matlab), Matweb (a TCP/IP client for the Matlab server) and the Matlab software tools.

The client, using a web browser, interacts with the e-learning data mining application by uploading the dataset in the server. The first step of the analysis is to request to the server the data descriptions. The result is the list of the columns' names that are loaded into a combo box.

The code in the web server is responsible for automatically constructing the html page that is sent to the client. In this way, the user can easily select the columns (e.g., columns for bi or multivariate analysis) that s/he wants to analyze. The interface supports a typical data mining workflow.

For univariate analysis, the interface offers a series of check boxes representing different types of indices that can be used to describe a selected variable: position indices (mean, etc), variability indices (difference between max and min etc), heterogeneity indices (entropy index, etc), concentration indices (Gini concentration index, etc), asymmetric index (skewness, etc), Kurtosis indices (Kurtosis index), hypothesis tests among the different modalities of the variable (ANOVA, etc), contingency tables and so on.

A similar approach is followed for the bivariate and multivariate analysis. In this case a server side code is responsible for constructing an html page containing the figures resulting from the analysis, e.g. a dispersion matrix.

Also, the tool supports a model construction. The client web interface is dynamically built containing the main parameters that can be chosen for the model, e.g. the number of nodes or the transfer function of the neural networks.

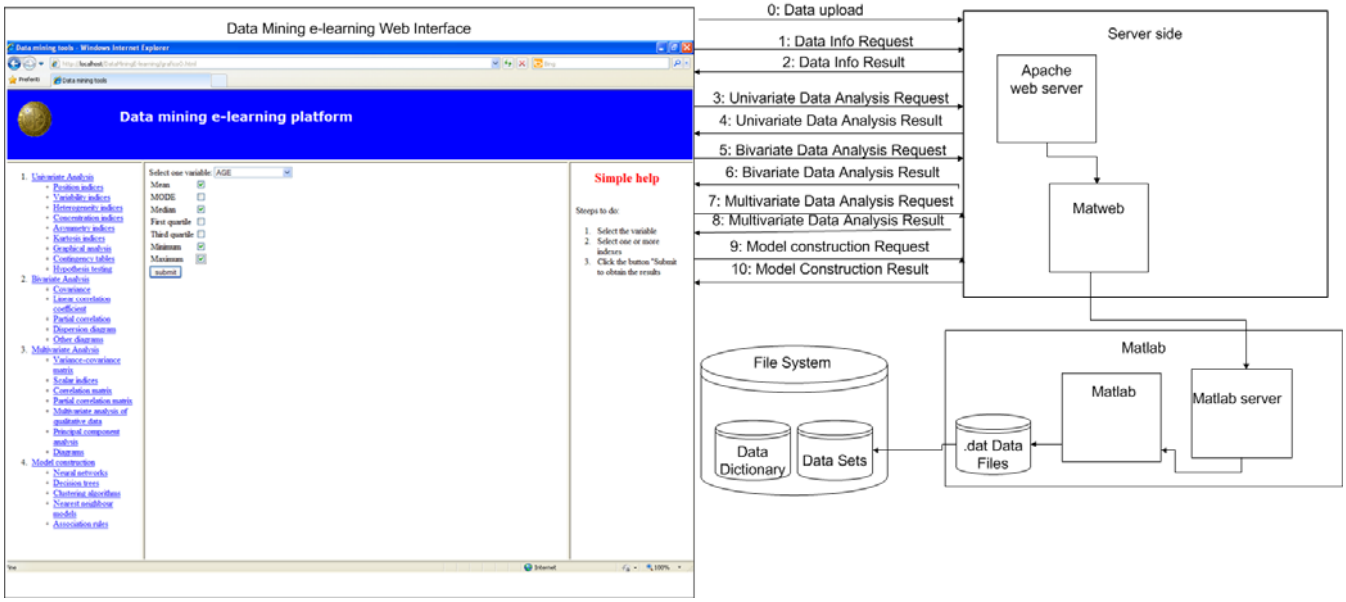

Fig.1. The data mining e-learning platform architecture.

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

All the datasets can be indexed so that they give rise to an organizational memory [12] that has been clustered in order to generate conceptual or thematic areas [13] in which the learners may find datasets similar to the one of their interest.

III. A Case Study

As a case study of a typical data mining analysis we summarize the experience of a group of students in an international data mining competition [9]. The case study is reported due to the extensive use of data analysis tools and techniques that can greatly benefit from the exploratory capabilities of the e-learning platform.

The students had to construct a model to predict, on the basis of customer usage and demographic data, which customers are likely to switch from 2G to 3G network. The data set consisted of 18,000 labeled rows: 15,000 belonging to 2G and 3,000 belonging to 3G with 251 columns: 214 quantitative and 37 qualitative. The analysis was performed separately for quantitative and qualitative columns.

For quantitative columns a first data cleaning phase removed the columns composed of all zeros or all NaN (Not a Number). Furthermore, the data whose values were outside the minimum and maximum range, specified in the data dictionary, were replaced by the mean of the remaining column values inside the class (2G or 3G).

The same replacement was performed for outliers identified by values more than 3σ away from the mean inside the category (2G or 3G). Columns composed of constant values (minimum equal to maximum) were removed. A min-max normalization step normalized all the variables in the range 0.1 – 0.9.

An ANOVA analysis was performed for all the remaining columns to test the null hypothesis of equal mean for the two groups (2G or 3G), therefore the columns for which the null hypothesis was not rejected were eliminated.

Similarly, by computing the linear correlation coefficient, one of the couple of columns, for which the null hypothesis of correlation was not rejected, was removed.

The previous analysis was deepened by computing the partial correlation coefficients among pairs of variables and then by removing one variable of the pair with a partial correlation coefficient higher than 0.95.

Moreover, the columns were discretized in eight bins of size 0.1: a contingency table for each variable composed of two rows (2G, 3G) and eight columns (one for each bin) was constructed. Table 1 reports an example of a contingency table for the first discretized variable.

Table I: An example of a contingency table

	C1	C2	C3	C4	C5	C6	C7	C8
3G	0,01	0,04	0,04	0,03	0,03	0,01	0,01	0,00
2G	0,07	0,18	0,19	0,15	0,14	0,07	0,03	0,01

An independency test with the Pearson statistics was performed allowing the removal of the columns for which the null hypothesis was rejected since the value distribution among the two class (2G, 3G) was similar.

Variables which were too homogeneous, or too heterogeneous, both for 2G and 3G were removed. Hence variables with a normalized Gini index or with an entropy index lower than 0.1, or higher than 0.9, both for 2G and 3G, were removed.

Finally a principal component analysis was carried out which may then lead to consider only the variables able to explain up to 95% of the overall variability.

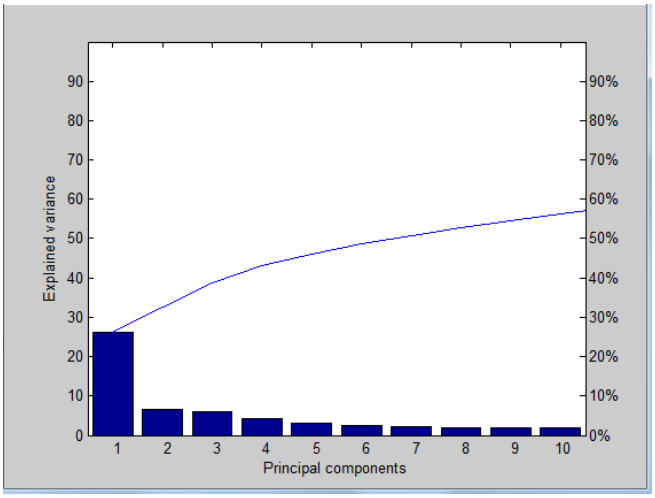

Fig. 2: The Pareto diagram for the first ten principal components.

Figure 2 shows the Pareto diagram for the explained variance. By analyzing the coefficients in the first principal component we recall some of the most influencing variables: the total number of different 1900 numbers called, average number of used days, the number of days since the last received retention campaign and the total number of different country codes called.

The overall data analysis process reduced the quantitative variables from 214 to 84. Of the qualitative columns, 20 were Boolean variables, 13 multiple codes and 4 variables with different modalities, with a maximum of 4 modalities.

The Boolean variables were binarized. With multi-modality variables a chi-square independency test on the contingency table was performed in order to remove variables with dependent values in the two modalities (2G, 3G). Similarly, variables which were too homogenous or heterogeneous were removed.

For variables with too many modalities, a clustering phase of their values was performed in order to reduce the number of variables created by the binarization process. At the end of the process 11 qualitative variables were selected which, after binarization, accounted for 26 input variables. The total number of variables that were used for the model construction was 110.

The available data were divided into learning and testing with a balanced number of samples from 2G and 3G. Trial were also done using a proportion of samples from 2G and 3G, as in the original dataset. The order of the input rows after the sampling procedure were randomized.

The final step of a typical data mining process consisted in a model construction phase. Different models, such as neural networks, were proposed.

For each model different configurations were studied and

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

standard techniques were used for their evaluation. For example, a multilayer perceptron neural network with 110 input neurons, two layers and one Boolean output neuron with a logsig transferring function was used. Several attempts were performed to fine tune other network parameters with a detailed comparison of the results. An iterative process was followed to establish the best number of neurons in the intermediate layer with different training epochs.

A team composed of two students from the University of Catania obtained a good performance in the PAKDD competition classifying in the upper half of the final ranking.

In the proposed scenario the e-learning platform can be valuable for data exploration purposes since it avoided the coding phase and allowed the students to concentrate on choosing the correct analysis and in result interpretation. For example, the user can choose the variable(s) to analyze, the indices and type of analysis to perform and finally, s/he has access to the results on a web page.

The modularity of the web platform allows one to extend and customize it to the growing needs of a data mining study and represents a small step in the Matlab world towards what the Weka platform [20] represents in the Java world.

IV. Conclusions

In this work we have highlighted the software architecture of an in house built e-learning environment for data mining analysis. We presented the software architecture and described a typical workflow with the e-learning application.

A case study of a complete dataset analysis is presented. The overall rating of the prototype of the data-mining e-learning platform was very positive since the platform offers tools that are valuable for students, especially in the stage of first analysis, by providing a fast exploratory medium.

In the future we plan to extend the algorithms and models supported by the application as well as planning a full scale evaluation experience and integrating other development tools to build server side .NET or Java components.

Also, we plan to link the datasets to the exercises done by the students, to ask the students to briefly describe their exercises with some screenshots, and to cite the exercises that have inspired their work.

This will generate another organization memory dealing with learners exercises where the students could find out how the datasets they choose have been mined by other learners in order to reuse experience by taking advantage from the textual and graphical annotations inserted by the previous learners [14], [15], [16].

Semantic description of both the datasets and exercises will favor the recall of the relevant data [17]. In particular we are experimenting how ontologies dealing with urban systems (e.g., [18], [19]) help learners in discovering with accuracy datasets from which they may derive rules for better managing urban processes (e.g., traffic, pollution, etc.) of their interest.

References

[1] The Joint Task Force on Computing Curricula Association for Computing Machinery, IEEE-Computer Society, "Computer Science Curricula 2013," available on line at http://ai.stanford.edu/users/sahami/CS2013/

[2] R. Mikut, M. Reischl, "Data Mining Tools," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol.1, no. 5, pp. 431-443, 2011.

[3] A. Faro, D. Giordano, F. Maiorana, "Mining Massive Datasets by an Unsupervised Parallel Clustering on a GRID: Novel Algorithms and Case Study," *Future Generation Computer Systems*, vol. 27, no. 6, pp. 711-724, 2011 .

[4] F. Maiorana, G. Fazio, "Knowledge Discovery from Text on a Cloud Architecture and its Application to Bioinformatics," in *Proc. 9th International Conference on Biomedical Engineering, IASTED* , 2012.

[5] A. Faro, D. Giordano, F. Maiorana, C. Spampinato, C, "Discovering Genes, Diseases Associations from Specialized Literature Using the GRID," *IEEE Transactions on Information Technology in Biomedicine*, Vol.13, no. 4, pp. 554-560, 2008.

[6] D. Giordano, F. Maiorana, "A Visual Tool for Mining Macroeconomics Data," *Management Information Systems*, vol. 10, pp. 241-251, WitPress, 2004.

[7] A. Faro, D. Giordano, F. Maiorana, "Input Noise Robustness and Sensitivity Analysis to Improve Large Datasets Clustering by Using the GRID". *Discovery Science, Lecture Notes in Computer Science* vol.5255, pp. 234-245: Springer Berlin/Heidelberg, 2008.

[8] S.M. Keaveney, M., Parthasarathy, "Customer Switching Behavior in Online Services: An Exploratory Study of the Role of Selected Attitudinal, Behavioral, and Demographic Factors," *Journal of the Academy of Marketing Science*, vol.29, no. 4, pp. 374-390, 2001.

[9] N.B. Noriel, C.L. Tan, "A Look Back at the PAKDD Data Mining Competition 2006," *International Journal of Data Warehousing and Mining*, vol.3, no. 2, pp. 1—11, 2007.

[10] P. Giudici, S. Figini, *Applied Data Mining for Business and Industry*", 2nd Edition, John Wiley & Sons, 2009

[11] J. Han, M. Kamber, *Data Mining: Concepts and Techniques*, 2nd ed. The Morgan Kaufmann Series in Data Management Systems, Jim Gray, Series Editor, 2006.

[12] A. Faro, D. Giordano, "Concept Formation from Design Cases: Why Reusing Experience and Why Not," *Knowledge Based Systems Journal*, vol.11, no. 7, pp. 437-448. Elsevier, 1998.

[13] A. Faro, D. Giordano, "Design memories as evolutionary systems: socio-technical architecture and genetics," *IEEE Proc. Int. Conf. on Systems, Man and Cybernetics*, Washington, D.C. USA., vol.5, pp. 4288-4293, IEEE, 2003.

[14] D. Giordano, "Evolution of interactive graphical representations into a design language: a distributed cognition account," *International Journal of Human-Computer Studies* Vol. 57, no. 4, pp. 317-345, 2002.

[15] S. Ahmed, "Encouraging reuse of design knowledge: a method to index knowledge," *Design Studies*, vol. 26, no. 6, pp. 565-592, 2005.

[16] A. Faro, D. Giordano, "StoryNet : an Evolving Network of Cases to Learn Information Systems Design," *IEEE Proceedings SOFTWARE*, vol.145, no. 4, pp. 119-127, 1998.

[17] A. Faro, D. Giordano, "Ontology, esthetics and creativity at the crossroad in information systems design," *Knowledge-Based Systems*, vol.13, no. 7, pp. 515-525, Elsevier, 2000.

[18] J. Teller, "Ontologies for an Improved Communication in Urban Development Projects," *Studies in Comp. Intelligence*, vol. 61, pp. 1-14, 2007.

[19] A. Faro, D. Giordano, A. Musarra, "Ontology based intelligent mobility systems," *IEEE Proc. Int. Conf. on Systems, Man and Cybernetics*, Washington, D.C. USA., vol.5, pp. 4288-4293. IEEE, 2003.

[20] I.H. Witten, E. Frank, Data Mining: Practical Machine Learning Tools and Techniques,3rd ed. Morgan Kaufmann Publishers, San Francisco (2011).

ISBN: 978-988-19251-3-8
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)