ISBN 978-83-60810-51-4

A Teaching Experience on a Data Mining Module

Francesco Maiorana
Department of Electrical,
Electronic and Computer
Engineering, University of
Catania. Viale Andrea Doria, 6,
Catania, Italy
Email:
francesco.maiorana@dieei.unict.it

***Abstract*—Data mining is recognized as an important field where one has the possibility to become accustomed both with analysis techniques and methods and with a state of mind. By means of data mining it is possible to develop critical skills that are essential in today's information technology. We present our experience in teaching a data mining module, within an Information System course, centered around a few key aspects: a convergence of theoretical Information Systems aspects and computing skills through programming a complete data mining analysis in Matlab; a project centered learning experience; a sharing of resources that are commented on both by the teacher and by peers facilitating the flow of information and the development of critical skills; a guided inquiry process where the students, when needed, are guided through appropriate questions in the right direction; and finally special attention to requiring motivation of each decision and step undertaken. As a case study we present and summarize the experience performed by two groups of students in a data mining study aiming at predicting a liquidity crisis of companies..**

I. Introduction

DATA mining has become an important topic in modern curriculum. The subject is recognized both in specialized courses and in modules within courses. In this regard we recall, just to cite a few: the Association of Computing Machinery (ACM) and the IEEE-Computer Society; the ACM and Association for Information Systems (AIS) Curriculum Guidelines for Undergraduate Degree Programs in Information Systems [2] and the graduate curriculum and guidelines for Information systems [3]. The field is a vibrating one with new techniques and methods developed such as: development of parallel algorithms that can be applied to massive datasets and deployed either on a grid architecture [4] or on a cloud architecture [5]; applying data mining techniques to literature mining problems allowing us to automatically extract new associations from scholarship documents [6] and estimating the impact of input noise on the sparse datasets [7] that derive from a vector space representation of the documents; techniques such as [8] to visualize data and results of the data mining procedure. The importance of data mining is broadened in [9] where the author suggests that 'data mining can serve as a tool for furthering information literacy' [10]. Vibrant is also the community around the teaching experience of data mining both at an Undergraduate and at a Master level. Among the work we would like to recall [11] where the author reports the teaching experience on a data mining course where the main objective of the course was to 'engage students in experimental research' and this objective was pursued through assignments, the reading and discussion of two published research papers and a project. In [12] the author also based his teaching approach, in a senior level data mining methods class, on practical projects and hands-on labs by using Microsoft Excell's data mining add-ins as a front-end and Microsoft's Cloud Computing and SQL Server 2008 Business Intelligence platforms as back-ends. In [13] the author presents an empirical evaluation of a data mining course reporting how the expertise of working students improved the course. In [14] the author also reports an experience focused on reading research papers and developing a project. In [15] the author presents the experience in teaching a data mining course offered for the first time. In [16] the author reports the difficulties in delivering data mining courses and designing well-structured curriculum in an environment that encompasses students with different backgrounds and knowledge. In [17] the author reports the experience in designing a curriculum for an elective data mining course in a graduate course in Information Systems. In [18] the author reports an experience in an undergraduate data mining course.

In this paper we report a four year teaching experience on a Master course in Information System where we taught a 30 hour module on data mining techniques and concepts. We present the course structure and organization and report and compare the experience in the various course editions. The approach used was a focus on Information Systems theory and practice in conjunction with computer science skills such as developing design and implementation capabilities of algorithms or ad hoc data-mining analysis in a high level and abstract language and framework such as Matlab. Data mining offers an excellent ground for developing critical thinking, creativity and problem solving and serves as a glue between Information System (IS) and Computer Science (CS) since its foundation heavily relies on both disciplines. The paper is organized as follows. Section II presents the course with particular emphasis on the data mining module highlighting the module objectives and project organization and assessment. Section III presents a brief overview of the range of problems studied over the years and, briefly, presents and discusses data mining studies aiming at predict-

978-83-60810-51-4/$25.00 © 2012 IEEE

ing the liquidity crisis of companies discussing the difference between two projects developed in different years. Section IV draws the conclusions and highlights future work.

II. The course structure

The course was taught at a Master level. The Information System course was a 60 hour course for a total of 6 credits. Half of the course dealt with Information System concepts and 30 hours dealt with data mining topics.

More than 50 students enrolled in the course each year with an almost constant number of students through the years. The majority of students were male but female students were well represented. The majority of students started the course with no or minimal experience in programming. Some of them during the course of the year were employed in local business companies.

The course objective with particular emphasis on the data mining module were: develop critical skills and capacity to analyze data from different perspectives, compare the analyses and draw conclusions; improve communication and cooperation skills; acquire the basic techniques and skills proper for data mining with the capacity to work out the field of usage of each technique as well as its strengths and drawbacks; acquire the capacity to use and program data mining tools and algorithms; apply techniques and tools in a multiple context data analysis scenario.

The course was built around a set of core concepts developed within in class sessions with small assignments requiring the application of the concept exposed in class. The small assignments were not graded. The argument of each lesson was exposed in advance and it was suggested to the student to read the material before the lessons. The syllabus of the course with its temporal division is presented in table 1 and [19] and [20] were used as reference books.

After a review of data mining tools available, both free and commercial, we chose to use the Matlab framework and its toolboxes. The suite offers a rich set of tools that can be extended with new algorithms and offers a way to develop applications as well as new algorithms. The suite was also used in a parallel course in statistics in the same semester. For a description of the suite and its usage in a web based learning tool the reader can reference [21]. The Matlab suite offers a good trade off by offering great flexibility and a rich set of instruments with the possibility to implement new algorithms or modifying existing ones. Matlab offers the basic programming structure such as assignments, conditional instruction, loops, function, array and cell array, structures and files that can be used to present the basic sequential programming construct. However, the great benefit in using Matlab derives from its high level structure that can be used to solve complex practical problems without requiring extensive coding while allowing us to concentrate on the most demanding and difficult challenge, i.e. designing the algorithm and developing adequate critical thinking and problem solving. The choice allows us to support the students in the implementation phase of the project by offering a rich set of tools that allows us to minimize the programming effort and concentrate on the analysis and problem solving effort that is the most difficult and important. There are studies, such as [22], that describe how Matlab was used in teaching programming skills where the author highlights how its use for introductory programming means avoiding losing the initial experience in a traditional programming language such as C, C++ or Java since these languages do not allow to be easily used for higher level technical problems in further courses. To ease even more the programming effort especially in undergraduate courses it is possible to use, in the initial phase, worked out examples with different settings and approaches such as have been described in [23]. At the end of the course a project was assigned to the students that were free to work in groups of up to three persons. The project was central for the module aiming at increasing critical skills by practically applying the course material requiring taking motivated decision by applying and comparing different techniques, working in groups and increasing communication and cooperation skills.

One of the key aspects of the project was to develop programming skills typical of a computer science course in programming through the design and development of a complete data mining analysis by programming a suite of scripts, one for each major phase of the analysis, in the Matlab environment. The project management was somewhat difficult, requiring extra work for the tutor. The tutor chose a rich set of initial data to work with and presented each dataset, the problem and same reference literature to the class, highlighting problems and basic techniques. The field of applications ranged from predicting customer behavior, predicting a liquidity crisis of companies, network intrusion detection, text mining applications such as mail spam detection, protein structure and biomedical signal classification, just to cite a few. After the initial presentation the student self-arranged into groups. Working in group was greatly encouraged. Duties and responsibilities were self-assigned by the group components and verified during the tutoring sessions and the final project presentation. The datasets were chosen from datasets made available in international competitions. The complexity of the dataset was chosen in such a way to be not too small and neither too large; a typical dimension is on the order of ten thousand rows and one hundred columns. The project required 4 to 6 weeks work. The tutor was available for meetings whenever requested by the students. The experience in the tutoring sessions reveals two broad categories of students: the most brilliant ones are able to work alone with minimal guidance and support only in critical phases. In this case the best interventions is in the direction of broadening the horizon and suggesting further work.

The second group of student requires more guidance that increases with decreasing levels of autonomy. As a result of the teaching experience care must be taken to avoid both a no intervention approach and a solution offering approach. The best strategy, but the most demanding one, is to pursue a guided inquiry approach trying to pinpoint errors and guide the students to self-discovery of solutions and correct approaches. This has also the advantage of increasing the self-confidence of the student that is more motivated in pursuing the project. In each intermediate deliverable, care should be used in requiring the reasoning behind and expla-

TABLE I.
COURSE SYLLABUS AND TEMPORAL SUBDIVISION

Topic	# of hours
Matlab tutorial	5
Exploratory data analysis: univariate	2
Exploratory data analysis: bivariate	2
Exploratory data analysis: multivariate	2
Distance and similarity measures	2
Clustering algorithms	2
Decision trees	2
Neural networks	3
Support vector machines	3
Nearest neighbor models	3
Methods to evaluate DM results	3
Projects presentations	1

nation of each decision using objective criterion whenever possible reinforced by research papers or previous studies. A mandatory session required students to present the final project deliverable and a power point presentation that will be used by the student in the final oral examination. This material was pre assessed by the tutor that required further intervention when necessary.

The module assessment was done with a written examination and by an oral presentation of the projects deliverable as well as the Matlab code. This evaluation was centered on the methodology used, not on the result obtained in terms of model accuracy. In the final presentation the students have to report the project milestone, document and motivate each decision taken, and make a comparison of two or more alternative solutions. A boost in the quality of the final project came from the adoption in the course of the year of a shared memory of projects [24]. The students had to look at the pool of projects, choose one or more project with either a similar case study or with similar techniques that the group plan to use. The chosen project must be peer reviewed and critically assessed by the group. The final deliverable must clearly state the proposed innovation with a clear motivation of the reason of improvements. The projects have clustered thematic areas around several dimensions such as the problem at hand or the data mining techniques used [25] in order to allow easier selection. Some projects were also presented and discussed in class by the tutor. Using a shared memory design is even more important and useful in data mining due to the increased level of uncertainty and the greater number of decision that must be taken.

III. A Case Study

As a case study we compare the project performed by two groups of student in different years with the same dataset.

The dataset used was the dataset publicly available in the 2005 German Classification Society (GfKl) Data Mining Competition aiming at predicting the possible liquidity crisis of companies. The training dataset was composed of 20.000 rows and 26 columns with a class distribution of 10% positive cases, i.e. where a liquidity crisis occurred, and 90% negative cases; the test data file was composed of 10.000 rows. The semantics of the attributes were not disclosed. Moreover, the dataset was affected by missing values, outliers and noisy distribution as often happens in real case scenarios. Students have performed a preliminary analysis and treatment of missing values which represents a difficulty in some types of dataset such as the one discussed here. For a recent work on treatment of missing values the reader can reference [26]. The unbalance of positive and negative cases requires attention and care must be taken to avoid overfitting. The first attempt with the dataset was performed in the 2004 – 2005 academic year. The students performed a good data analysis, obtaining univariate and bivariate statistics, a treatment of zero and 9.999.999 values, outlier detection and treatment, categorical variable analysis, binarization and dimensionality reduction and so on. In the project each decision was clearly motivated. A second group of students in the 2005 – 2006 academic year, using the shared memory of previous case, chose the same project. As a result of the peer review of the previous case study students analyzed the positive aspects as well as the negative aspects. Their work started from the previous experience and as a result the final project was much better: the data analysis and the cleaning and transformation phases were done with the same care but by using different techniques exploiting ad-hoc and self-invented graphs. The model construction part was done with greater care, with a detailed analysis and fine tuning of model parameters and by comparing the results of three different techniques such as neural networks, decision trees and support vector machines. This represents a clear broadening and deepening of the model construction phase compared with the first project where only a neural network model analyzed with few network configurations and parameters was developed. The student reported the great help and increased confidence obtained in using and reviewing the previous experience.

IV. Conclusions

In this work we have presented a four years teaching experience in a data mining module inside a Master course in Information Systems. The experience was centered around a project. Among the key points it is possible to recall: a practical group project that serves as glue between theory and practice and increases communication and collaboration skills as well as critical thinking and problem solving skills; use of the Matlab suite to design and implement, by means of a powerful high level language, a complex data mining project; a shared memory of project cases, peer reviewed by the students and assessed by the instructors facilitating the flow of information and the development of critical skills; a guided inquiry process; special attention in requiring motivating each decision and step undertaken. In the future 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 chose 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 [27], [28], [29]. Semantic description of both the datasets and exercises will favor the recall of the relevant data [30]. In particular we are experimenting how ontologies dealing with urban systems (e.g., [31], [32]) help learners in discovering with accuracy datasets from which they may derive rules for better managing urban processes 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] H. Topi, J. S. Valacich, R. T. Wright, K. M. Kaiser, J. F. Nunamaker, Jr., J. C. Sipior, G.J. de Vreede, "IS 2010 Curriculum Guidelines for Undergraduate Degree Programs in Information Systems," Association for Computing Machinery (ACM) Association for Information Systems (AIS), 2010 avaiable on line at http://www.acm.org/education/curricula/IS%202010%20ACM%20final.pdf

[3] J. T. Gorgone, P. Gray, E. A. Stohr, J. S. Valacich, R. T. Wigand, "MSIS 2006: model curriculum and guidelines for graduate degree programs in information systems, " *Communications of AIS, Volume 17, Article 1*

[4] 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 .

[5] 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.

[6] 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.

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

[9] B. Berendt, "Data mining for information literacy", *Data Mining: Foundations and Intelligent Paradigms*, Dawn E. Holmes and Lakhmi C. Jain. (Eds.), Springer 2011.

[10] J. J. Shapiro, S. K. Hughes, "Information literacy as a libera art.

[1] enlightenment proposals for a new curriculum," Educom Review, 31, no. 2, 1996.

[11] I. Rahal, "Undergraduate research experiences in data mining"," in in Proc. *39th SIGCSE Technical Symposium on Computer Science Education,* ACM, 461-4, 2008.

[12] M. J. Jafar, "A Tools-Based Approach to Teaching Data Mining Methods," Journal of Information Technology Education: Innovations in Practice, Vol. 9, 2010.

[13] P. Christen, "Evaluation of a graduate level data mining course with industry participants"

[14] D. R. Musicant, "A data mining course for computer science: primary sources and implementations", in Proc. *37th SIGCSE Technical Symposium on Computer Science Education,* ACM, 461-4, 2006.

[15] N. V. Chawla, "Teaching data mining by coalescing theory and applications," in Proc. ASEE/IEEE frontiers in education conference, Indianapolis, 2005

[16] B. Stewart, "Reflection on development and delivery of a Data Mining unit," Proceeding of Sixth Australian Data Mining Conference, Gold Coast, Australia, 2007. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 70. Peter Christen, Paul Kennedy,Jiuyong Li, Inna Kolyshkina and Graham Williams, Ed.

[17] S.R. Ponelis, "Finding Diamonds in Data: Reflections on Teaching Data Mining from the Coal Face," Issues in Informing Science and Information Technology, vol.6, pp. 227-240, 2009

[18] C. Schmidt, "Lesson learned in the design of an undergraduate data mining course," Journal of Computing Sciences in Colleges, vol.26, no.5, ++. 189-195, 2011.

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

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

[21] F. Maiorana, A. Mongioj, M. Vaccalluzzo, "A data mining E-learning tool: description and case study," Proceedings of the 2012 International Conference of Data Mining and Knowledge Engineering, London, U.K., 4-6 July, 2012.

[22] M. E. Herniter, D. R. Scott, "Teaching programming skills with Matlab," Proccedings of the 2001 American Society for Engineering Education annual conference and exposition,., 2001

[23] I. C. Moura, "Worked-out Examples in a Computer Science Introductory Module," Proceedings of the World Congress on Engineering 2012 Vol II, WCE 2012, July 4 - 6, 2012, London, U.K.

[24] 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.

[25] 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.

[26] M. M. Rahman, D. N. Davis, "Fuzzy Unordered Rules Induction Algorithm Used as Missing Value Imputation Methods for K-Mean Clustering on Real Cardiovascular Data," Proceedings of the World Congress on Engineering 2012 Vol I WCE 2012, London, U.K.

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

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

[29] 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.

[30] 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.

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

[32] 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.