Second International Conference
Modelling and Development of Intelligent Systems
Sibiu - Romania, September 29 - October 02, 2011

Versatile integration of data mining techniques of description and prediction in Web informatics systems of Business Intelligence

Mircea – Adrian MUSAN

Abstract

Using data mining techniques in computer applications for the digital economy, but not only, opened new possibilities for handling information in real time and their movement between all the factors involved. Computer applications based on these techniques assist successfully entrepreneurs in making decisions to achieve a higher degree of economic efficiency at company / organization level. Through this paper Web framework proposed, presented and developed is a versatile application, designed in a flexible way to integrate data mining techniques that are represented by RapidMiner processes applied in e-business. My application proposed here is in fact a management system of users and its processes assigned, build by data mining techniques, preprogrammed in RapidMiner.

1 Introduction

Application proposed here through this work is a departure point, one component of the subsequent developments, which can be used in the construction of Web informatics systems with application in economic areas, but not only.

1.1 Technology for extraction and knowledge management: data mining

Data mining techniques provides a useful and broad perspective in developing and using information systems in the field of e-business. Diverse areas and purposes of use, and the need to remote access, are elements of departure in this work. Thus, by subject matter, it wants to join the existing international concerns in the field of information technologies and their integration into economic applications.

Data mining is the process of extracting the knowledge on the databases / data warehouses, knowledge previously unknown, valid and operational at the same time [1].

Data mining, the analysis step of the Knowledge Discovery in Databases process, or KDD [5] is the process of extracting patterns from large data sets, by combining methods from statistics and artificial intelligence with database management. [6] With recent tremendous technical advances in processing power, storage capacity, and inter-connectivity of computer technology, data mining is seen as an increasingly important tool by modern business, to transform unprecedented quantities of digital data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud

detection, and scientific discovery. The growing consensus that data mining can bring real value has led to an explosion in demand for novel data mining technologies. [7]

Data mining deals with large complex data processing. Robust tools are needed to recover weak signals. These instruments require extremely efficient algorithms to achieve the desired processing. Data mining software combines artificial intelligence, statistical analysis and systems management databases to try to extract knowledge from data stored.

1.2 RapidMiner, "engine" in development processes through data mining techniques

Rapid Miner is an environment for machine learning, data mining, text mining, predictive analytics, and business analytics. It is used for research, education, training, rapid prototyping, application development and industrial applications. [8]

It allows experiments to be made up of a large number of arbitrarily nestable operators, described in XML files which are created with RapidMiner's graphical user interface. It provides a GUI to design an analytical pipeline (the "operator tree" in RapidMiner parlance). The GUI generates an XML (eXtensible Markup Language) file that defines the analytical processes the user wishes to apply to the data. This file is then read by RapidMiner to run the analyses automatically. [2]

For my application I realized the construction of a series of processes RapidMiner (an open source environment that provides complex processing to achieve modular operators) by using data mining techniques specific to e-business, using techniques of description and prediction. A data mining process is viewed as a complex process which consists in carrying out a sequence of steps: data cleaning, integration of different sources, selecting relevant data, the summary data transformation and aggregation, knowledge extraction and presentation templates.

2 Building data mining processes as tools for the investigation of specific e-business applications

2.1 Data Mining Technologies

The tasks of data mining process can be classified by types of knowledge sought by the user. The most common types of data mining tasks are [2][3]:

- *Summarizing*: A set of relevant data is summarized and abstracted resulting in a smaller set of data providing an overview of the aggregated information. A summary table can be generalized to different levels of abstraction and seen from different angles. For example, a company can be summarized sales by product, region or years and seen various levels of abstraction in any combination thereof.

- *Classification*: In the process of classification is given to analyzing a training data set or a set of objects whose class label is unknown. The model can be used to classify future data and develop a better understanding of each class in the database. For example, a model can be built based on the classification of disease symptoms and characteristics that can be used to diagnose new patients.

- *Clustering*: Clustering is the process of site identification for a variety of classes of objects based on attributes unclassified them. Objects that are classified as inter-class similarities are minimized by relying on certain criteria. Once you have decided clusters with common features of objects in a cluster, they are summarized to form a description of the class. For example, a company can classify its customers into several categories based on similarities of their age, income or address, and the common characteristics of customers in a category can be used to describe that group of customers.

- *Analysis of trends*: Templates and regularities in the behavior of data changes are detected during data binding (which they attribute data time). The data are analyzed over time and trace data are compared and appropriate modification. Trends, such as periods of growth or decay times, which happens frequently reported. For example, sales of companies can be analyzed each year, quarterly or monthly sales and to discover models to analyze the reasons behind them.

- *Mining association rules based*: An association rule reveals the associative relationships between objects, especially in a transactional database. For example, an association rule, "expected message, a message appears", says if a customer subscribes to the service "call waiting" is very likely that he or she also have the service "call display". Databases are searched to identify associations between objects and data. Another example, a retail store may discover that a set of goods is often bought with a different set of goods. This finding can then be used to design the sales strategy.

2.2 Development of RapidMiner processes through data mining techniques

Trough my work I realized the construction of a series of processes data mining through RapidMiner by using data mining techniques of description and prediction. A data mining process is viewed as a complex process which consists in carrying out a sequence of steps: data cleaning, integration of different sources, selecting relevant data, the summary data transformation and aggregation, knowledge extraction and presentation templates.

Following the versatility – flexibility axis, the processes are developed from a wide area, using such classification and regression techniques, clustering and association analysis techniques. The techniques used in the construction processes of classification and regression methods are: k-NN, DecisionTree, SuportVectorMachine, clustering methods: K-Means and BDScan, by association analysis methods: W-Apriori and FPGrowth – Association Rules.

Before developing the application was needed to build a series of processes with data mining techniques, by programming in RapidMiner, using such predictive techniques but also description techniques.

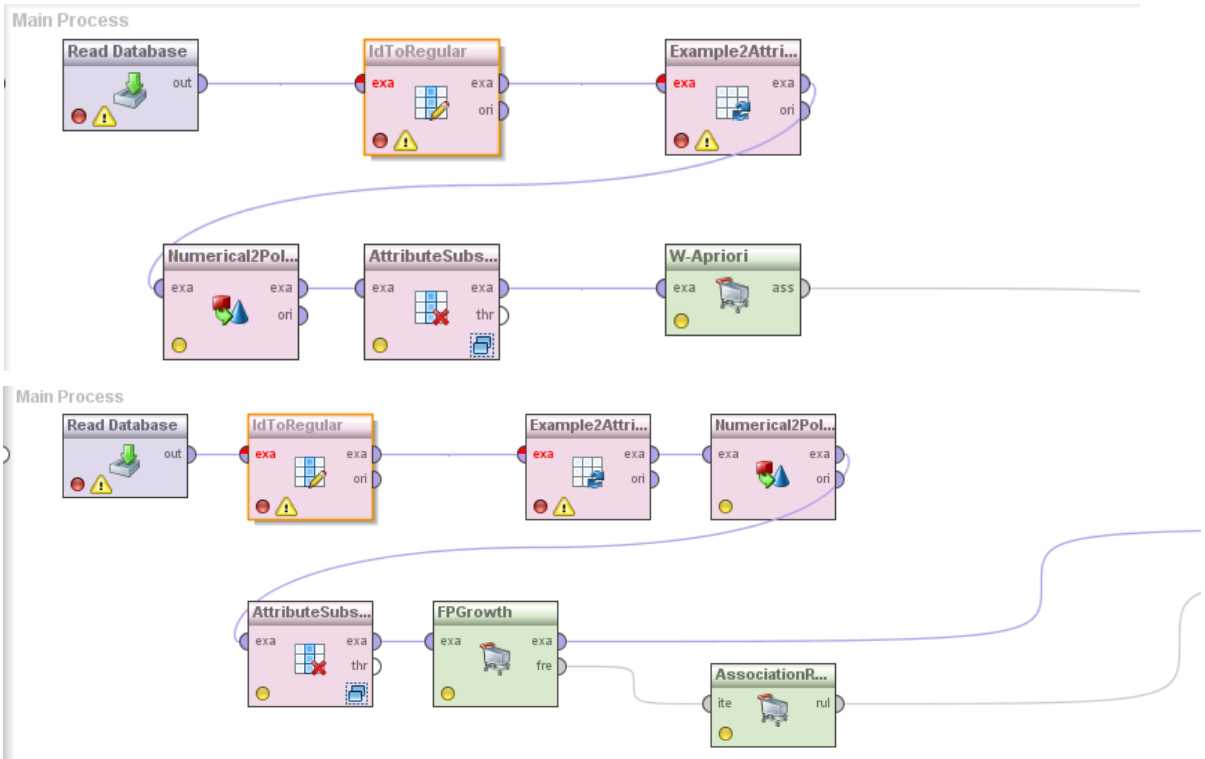

Figure 1 – Such examples built on description techniques

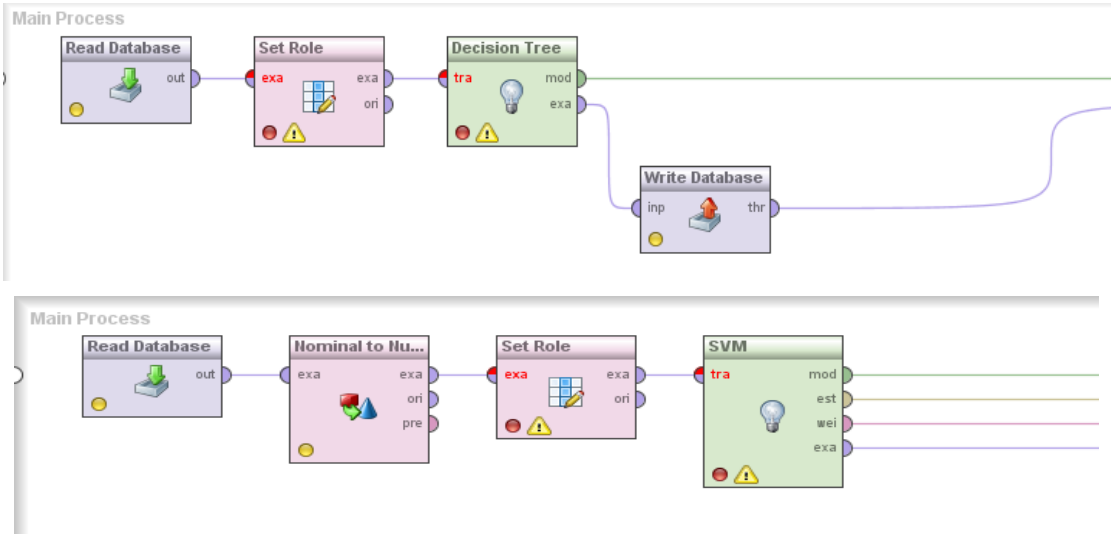

Figure 2 – Such examples built on prediction techniques

3 The Web application developed for integration of data mining generic techniques

3.1 The application structure

The application submitted through this work is created by two client – side components:

- application instantiator enables the instantiation of new data mining web applications and the registration of users for which the access to application resources is granted;

- application manager allows registered users to manage all the actions for which they granted permission, such as connecting to the data mining engine hosted by a server, connecting to remote databases, editing, running and modifying complex data mining experiments, visualizing the results, and so on. [4]

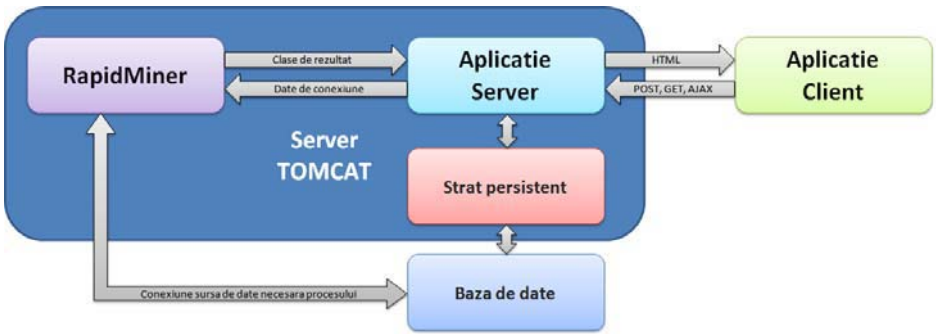

Figure 3 – The application structure

In Figure 3 we can see structural form of the application:

- *Persistent leyer* (*Hibernate*): intermediate trading of commands SQL and assist in handling data

- *RapidMiner*: compose from RapidMiner library it running and execute processes

- *Server application*: the answers transpose into HTML and sends it to client application.

3.2 Versatility – flexibility – generality axis conferred

The following features bring the application in versatility – flexibility – generality axis:

- Any user, according to his membership in a group, can load and run a desired process from the categories of processes integrated in the application.

- There was processed a number of data mining processes developed in RapidMiner and rendering of the results is specific for each particular type of process.

- The user can add new processes, which, if built on some classes of operators, after scanning, application recognizes respective process, can fit into a particular class and can send that result.

- There is the possibility of introducing of other types of processes, but, given the characteristic way of execution and the results of its own way of rendering, this can only be programmed into the application environment.

- Any existing process on the application server can be accessed through its XML code.

- There can be used different datasets as storage diversified through files in various formats (xls,. arff, .aml, .csv, .sql, c4.5, .dat, .xrff, etc..). Any set of data may be retrieved from a host computer, from a URL, from server application's database or from another database server, and then it is converted to an SQL database and added to current server database.

- It can use the data sets diversified in content, provided by a structure that would be acceptable to the process.

3.3 Presentation of the application

The developed framework is an easy management users and their associated processes system, for analysing data and making decision, designed to run any data mining process, by using the algorithms from the specific e-business application classes. Through its architecture, the attached application provides a highly flexibility, can easily be modified in content, distributed and improved, as a new element in advanced use of intelligent systems in economic activities. Such an application can be regarded as a flexible informatics system, due to its techniques and due to its heterogeneous data sets used for work, as well as to its accessing mode, for anybody and from anywhere, through the "browser" tool used for integrating, adding, accessing and view of the results rendered particularly for each implemented technique, and the possibility of modifying, and deleting the realised processes, respectively.

It is presented in detail how to use the application, entry window, with options for user's access, management options for user groups, data mining process management option, the option of running processes and logout option. For each of these options, the content panel displays the controls for execution and views the results.

Figure 4 – The user's list

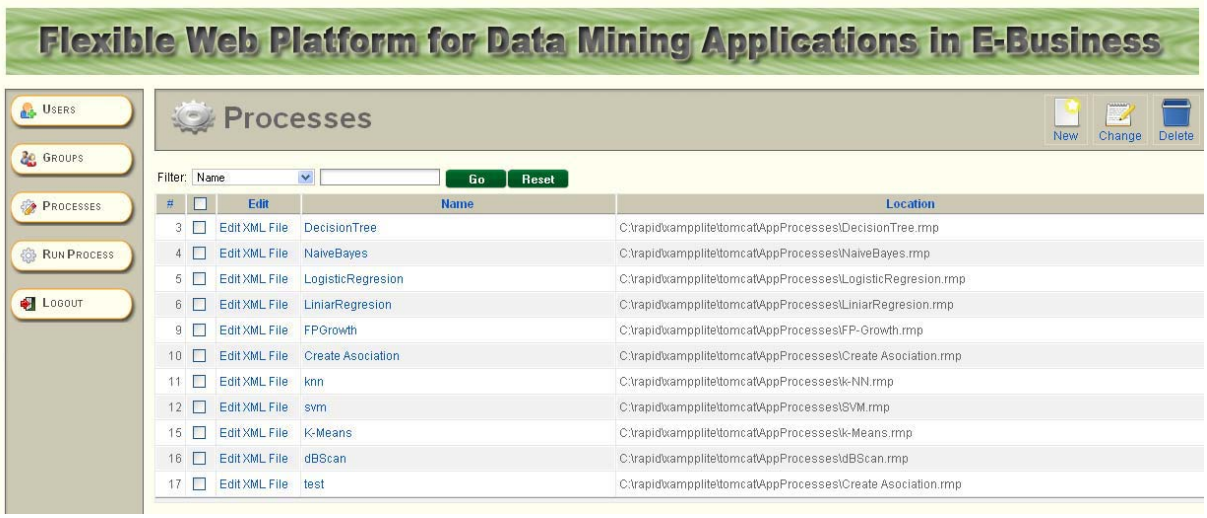

Figure 5 – Management of processes

Figure 6 – Possibility of modifying the existing processes wrote in XML code

Figure 7 – Running a RapidMiner process

Update dataset can be done by providing a link where the *.csv* file containing the new set of data or providing the address and authentication data on a server database and an order of selection set. Running process is done by clicking on the link associated *Run process*. This will generate a form dynamically depending on the current process.

The results of a process are stored in the database as HTML text, which is created dynamically according to each process. Processes can have one or more results, but they all share a schematic representation of the data upon which it was run.

I will present here viewing mode of the results for two any processes loaded on applications: one of prediction and one of description.

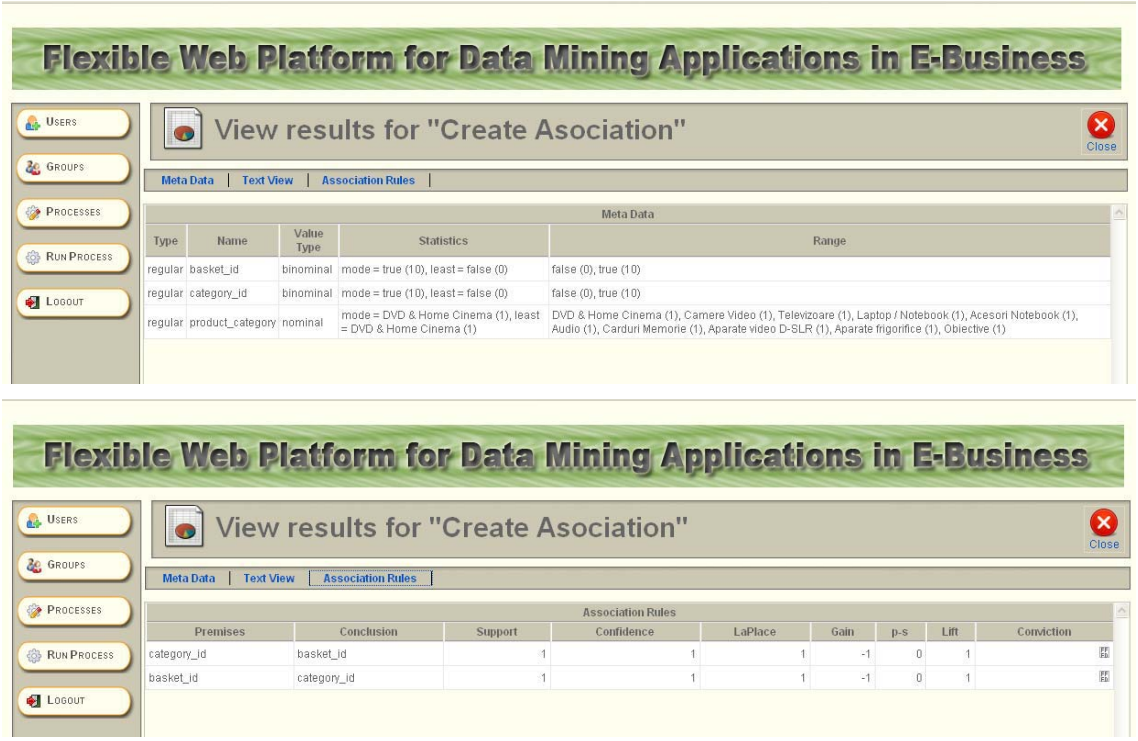

Figure 8 – Ways to view results for a particular process built by description techniques

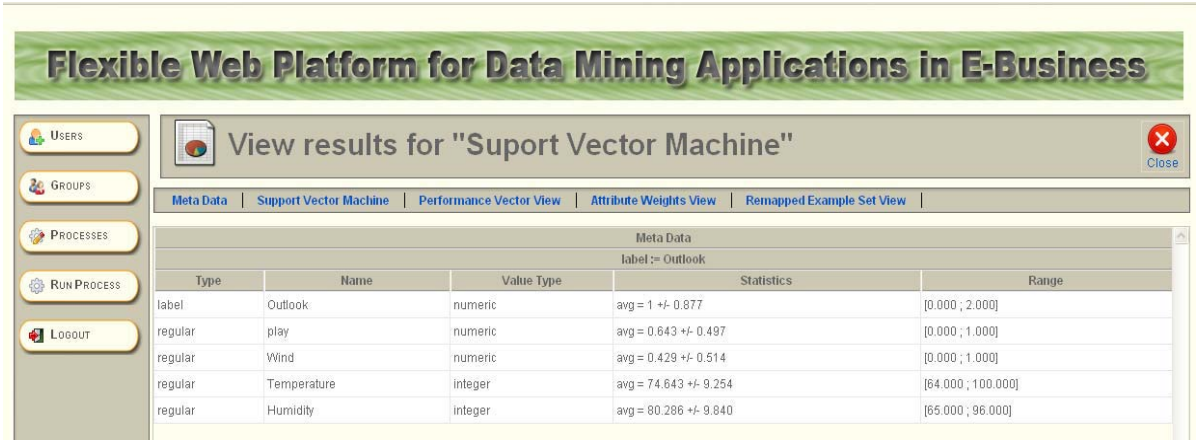

Figure 9 – Ways to view results for a particular process built by prediction techniques

As it can see each mode of rendering the results is particular for each technique, regardless of the class to which it belongs. All these modes of rendering the results are those proposed by RapidMiner at viewing its results and integrated by my application.

4 Conclusions

In terms of practical usefulness, such an application developed can be regarded as a system usable in e-business for a company with several departments (development, innovation, marketing, promotion, etc.), for each of them being possible combinations of processes built into RapidMiner, or it can be regarded as a Web platform that can be used by anyone, anywhere, for the integrating, adding and accessing data mining techniques build processes.
The originality of this application lies in the idea of building it, and in the implemented methodology, by developing of an application of process integration, built by using data mining techniques in a flexible and general mode, but with the possibility of assimilation to e-business systems.

References

[1] Giudici P., *Applied Data Mining: Statistical Methods for Business and Industry*, J. Wiley & Sons, 2005
[2] Mușan M., Hunyadi D., *Opportunities in development of business systems by building a web flexible framework for integrating generic data mining techniques in economic activities*, Analele Universității "Eftimie Murgu" Reșița, Fascicolul II Stiințe Economice, CNCSIS tip B+, pp. 380-389, 2010, ISSN 1584-0972 (CNCSIS B+)
[3] Hunyadi D., Mușan M., *Integration of data mining techniques in e-commerce applications*, Analele Universității "Eftimie Murgu" Reșița, Fascicolul II Stiințe Economice, CNCSIS tip B+, pp. 568-576, 2010, ISSN 1584-0972 (CNCSIS B+)
[4] Georgescu V., *WSRP-Enabled Distributed Data Mining Services Deliverable over a Knowledge-Driven Portal*, Proceedings of the 7th WSEAS International Conference on Applied Computer & Applied Computational Science (ACACOS'08), April 6-8, 2008, Hangzhou, China, 1790-5117, 150-156
[5] Christopher Clifton, *Encyclopedia Britannica: Definition of Data Mining*, http://www.britannica.com/EBchecked/topic/1056150/data-mining Retrieved 2010-12-09, 2010
[6] http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[7] http://www.sigkdd.org/curriculum.php
[8] http://en.wikipedia.org/wiki/RapidMiner

Mircea – Adrian MUSAN
"Lucian Blaga" University from Sibiu
Department of Mathematics and Informatics
Sibiu, Ion Ratiu Street, No. 5 – 7
Romania
E-mail: musanmircea@yahoo.com