The Role of Domain Knowledge in a Large Scale Data Mining Project

Ioannis Kopanas, Nikolaos M. Avouris, and Sophia Daskalaki

University of Patras, 26500 Rio Patras, Greece
(ikop, N.Avouris}@ee.upatras.gr, sdask@upatras.gr

Abstract. Data Mining techniques have been applied in many application areas. A Data Mining project has been often described as a process of automatic discovery of new knowledge from large amounts of data. However the role of the domain knowledge in this process and the forms that this can take, is an issue that has been given little attention so far. Based on our experience with a large scale Data Mining industrial project we present in this paper an outline of the role of domain knowledge in the various phases of the process. This project has led to the development of a decision support expert system for a major Telecommunications Operator. The data mining process is described in the paper as a continuous interaction between explicit domain knowledge, and knowledge that is discovered through the use of data mining algorithms. The role of the domain experts and data mining experts in this process is discussed. Examples from our case study are also provided.

1 Introduction

Knowledge discovery in large amounts of data (KDD), often referred as data min-ing, has been an area of active research and great advances during the last years. While many researchers consider KDD as a new field, many others identify in this field an evolution and transformation of the applied AI sector of expert systems or knowledge-based systems. Many ideas and techniques that have emerged from the realm of knowledge-based systems in the past are applicable in knowledge discovery projects. There are however considerable differences between the traditional knowledge-based systems and knowledge discovery approaches. The fact that today large amounts of data exist in most domains and that knowledge can be induced from these data using appropriate algorithms, brings in prominence the KDD techniques and facilitates the building of knowledge-based systems. According to Langley and Simon [1] data mining can provide increasing levels of automation in the knowledge engineering process, replacing much time-consuming human activity with automatic techniques that improve accuracy or efficiency by discovering and exploiting regularities in stored data. However the claim that data mining approaches eventually will automate the process

I.P. Vlahavas and C.D. Spyropoulos (Eds.): SETN 2002, LNAI 2308, pp. 288–299, 2002.

and lead to discovery of knowledge from data with little intervention or support of domain experts and domain knowledge is not always true.

The role of the domain experts in KDD projects has been given little attention so far. In contrary to old knowledge-based systems approaches where the key roles were those of the domain expert and the knowledge engineer, today there have been more disciplines involved that seem to play key roles (e.g. data base experts, data analysts, data warehouse developers etc.) with the consequence the domain experts to receive less prominence. Yet, as admitted in Brachman &Anand [2], the domain knowledge should lead the KDD process. Various researchers have made suggestions on the role of domain knowledge in KDD. Domingos [3] suggests use of domain knowledge as the most promising approach for constraining knowledge discovery and for avoiding the well-known problem of data overfitting by the discovered models. Yoon et al. [4], referring to the domain knowledge to be used in this context, propose the following classification: inter-field knowledge, which describes relationship among attributes, category domain knowledge that represents useful categories for the domains of the attributes and correlation domain knowledge that suggests correlations among attributes. In a similar manner Anand et al. [5] identify the following forms of domain knowledge: attribute relationship rules, hierarchical generalization trees and constraints. An example of the latter is the specification of degrees of confidence in the different sources of evidence. These approaches can be considered special cases of the ongoing research activity in knowledge modeling, ontologies and model-based knowledge acquisition, see for instance [6], [7], with special emphasis in cases of data-mining driven knowledge acquisition.

However these studies concentrate in the use of domain knowledge in the main phase of data mining, as discussed in the next section, while the role of domain knowledge in other phases of the knowledge discovery process is not covered. In this paper we attempt to explore our experience with a large-scale data-mining project, to identify the role of the domain knowledge in the various phases of the process. Through this presentation we try to demonstrate that a typical KDD project is mostly a multi-stage knowledge modelling experiment in which domain experts play a role as crucial as in any knowledge-based system building exercise.

2 Identification of Key Roles and Key Phases of a KDD Project

According to Langley and Simon [1] the following five stages are observed in the development of a knowledge-based system using inductive techniques: (a) problem formulation, (b) determination of the representation for both training data and knowledge to be learned, (c) collection of training data and knowledge induction, (d) evaluation of learned knowledge, (e) fielding of the knowledge base. In Fayyad et al. [8] the process of KDD is described through the following nine steps: (a) Defining the goal of the problem, (b) Creating a target dataset, (c) Data cleaning and preprocessing, (d) Data transformation, e.g. reduction and projection in order to obtain secondary features, (e) matching the goals of the project to appropriate data mining method (e.g. clustering, classification etc.), (f) Choosing the data mining algorithm to

be used, (g) Data Mining, (h) Interpretation of identified patterns, (i) Using discovered knowledge. By comparing the two processes one should notice the emphasis of the first frame on knowledge and the second on data analysis and processing. However in reality, while the stages proposed by Fayyad et al. do occur in most cases, the role of domain knowledge in them is also important, as discussed in this paper, while the final stage of building the knowledge base and fielding the system is also knowledge-intensive, often involving multiple knowledge representations and demanding many knowledge evaluation and knowledge visualization techniques.

The subject of our case study was the development of a knowledge-based decision support system for customer insolvency prediction in a large telecommunication industry. During the initial problem definition phase the observation of existence of large amounts of data in the industry concerned, led to the decision of extensive use of KDD techniques during this project. However this extensive dataset did not cover all aspects of the problem. While high degree of automation in modern telephone switching centres means that telephone usage by the customers of the company was well monitored, information on the customers financial situation and credit levels, which are particularly important for this problem, were missing. This is a problem that often occurs in real problems; that is different levels of automation in different aspects of the problem domain leads to non-uniform data sets. Also techniques to infer knowledge, based on assumptions, observations and existing data need to be used extensively during the problem definition and modeling phase. So, for instance, if information on the credit levels of a customer is missing, this can be inferred from information on regularity of payment of the telephone bills, based on the assumption that irregular payments are due to financial difficulties of the customers involved. This is a typical example of use of domain knowledge in the so called ‘data transformation phase’. From early stages it was deduced that a number of domain experts and sources of data had to be involved in the process. Domain experts, e.g. executives involved in tackling the problem of customer insolvency and salesmen who deal with the problem in day-by-day basis were interviewed during the problem formulation phase and their views on the problem and its main attributes were recorded. An investigation of the available data was also performed and this involved executives of the information systems and the corporate databases who could provide an early indication on sources and quality of data. Other key actors were data analysts, who were also involved together with knowledge engineers and data mining experts.

3 Business Knowledge and KDD

In many KDD projects, like the case study discussed here, the domain knowledge takes the form of business knowledge, as this represents the culture and rules of practice of the particular company that has requested the knowledge-based system. Business knowledge has been a subject of interest for management consultant firms and business administration researchers [9]. Business process re-engineering (BPR) is a keyword that has been extensively used during the last years, while special attention has been put in building the so-called “institutional memory”, “lessons learned data

bases" and "best practice repositories". While there are but few examples of successful full-scale repositories of business knowledge in large companies today, the widespread application of these techniques makes worth investigating their existence. The relevance of these approaches to KDD projects and the importance of them as sources of domain knowledge to data mining efforts is evident and for this reason they should be taken in consideration. It should also be noticed that a side effect of a major data-mining project could be the adaptation of a business knowledge base with many rules and practices, which resulted from the KDD process. This is also the case with tacit knowledge and implicit knowledge, which is the not documented business knowledge, often discovered during such a project.

The distinction between domain knowledge and business knowledge is that the former relates to a general domain while the latter to a specific business, thus both are required in the case of a specific knowledge based system that is to be commissioned to a specific company. A special case of business knowledge that affects the KDD process relates to the business objectives as they become explicit and relate to problem definition. These can influence the parameters of the problem and measures of performance, as discussed by Gur Ali and Wallace [10]. In the following an example of such mapping of business objectives to measures of system performance is described for our case study.

4 Use of Domain Knowledge in an Insolvency Prediction Case Study

In this section, some typical examples of applying domain and business knowledge in the case study of the customer insolvency problem are provided. The examples are presented according to their order of appearance in the different phases of the KDD process. In the following section a classification of the domain and business knowledge used is attempted. The discussion included in this section does not provide a full account of the knowledge acquisition and modeling case study. It attempts rather through examples to identify the role of domain knowledge in the various phases of the project. A more detailed account would have included details on the modelling process, which involved many iterations and revisions of discovered knowledge.

A detailed description of the problem of customer insolvency of the telecommunications industry is beyond the scope of this paper. For more information on the problem, the approach used and the performance of the developed system, see Daskalaki et al. [11].

4.1 Problem Definition

In this phase the problem faced by a telecommunications organization was defined and requirements relating to its solution were set. The role of domain experts and the importance of domain and business knowledge in this phase is evident. For instance the billing process of the company, the rules concerning overdue payments and currently

applied measures against insolvent customers need to be explicitly described by domain experts.

4.2 Creating Target Data Set

Formulation of the problem as a classification problem was performed at this stage. Available data were identified. As often occurs in KDD projects available data were not located in the same database, while discrepancies were observed among the entities of these databases. This phase was not focused on the specific features to be used as parameters for training data, but rather on broad data sets that were considered relevant, to be analyzed in subsequent steps. So the sources of data were: (a) telephone usage data, (b) financial transactions of customers with the company (billing, payments etc.), (c) customer details derived from contracts and phone directory entries (customer occupation, address etc.). As discussed earlier more details of customer credit conditions were not available in the corporate databases and could not become available from outside sources. The role of domain and business knowledge in this stage concerned the structure of the available information and the semantic value of it, so this knowledge was offered mostly by the data processing department, in particular employees involved in data entry for the information systems involved. Serious limitations of the available data were identified during this process. For instance it was discovered that the information systems of the organization did not make reference to the customer as an individual in recorded transactions, but rather as a phone number owner. This made identification of an individual as an owner of multiple telephone connections particularly difficult.

4.3 Data Prepossessing and Transformation

This phase is the most important preparation phase for the data mining experiments; The domain knowledge during this stage has been used in many ways:

(i) elimination of irrelevant attributes
(ii) inferring more abstract attributes from multiple primary values
(iii) determination of missing values
(iv) definition of the time scale of the observation periods,
(v) supporting data reduction by sampling and transaction elimination

In all the above cases the domain knowledge contributes to reduction of the search space and creation of a data set in which data mining of relevant patterns could be subsequently performed. Examples of usage of domain knowledge are:

Example of case i: The attribute “billed amount” was considered as irrelevant since it is known that not only insolvent customers relate to high bills, but also very good solvent customers.

Examples of case ii: Large fluctuation of the amounts in consecutive bills is considered important indication of insolvency, so these fluctuations should be esti-mated and taken in consideration.

Overdue payments have been inferred by comparison of due and payment dates of bills.

Considerable reduction of data was achieved by aggregating transactional data in the time dimension according to certain aggregation functions (sum, count, avg, stddev) and deduced attributes. Domain-related hypotheses of relevance of these deduced attrib-utes have driven this process. An example was the *DiffCount* attribute that represents the number of different telephone numbers called in a given period of time and the deviation of this attribute from a moving average in consecutive time periods. Defini-tion of this attribute is based on the assumption that if the diversity of called numbers fluctuates this is an entity related to possible insolvency.

Example of case iii: In many cases the missing values were deduced through interrelated attributes, e.g. Directory entries were correlated with customer records in order to determine the occupation of a customer, payments were related to billing periods, by checking the amount of the bill etc.

Examples of case iv: The transaction period under observation was set to 6 months prior to the unpaid bill, while the aggregation periods of phone call data was set to that of a fortnight.

Example of case v: Transactions related to inexpensive calls (charging less than 0.3 euros) were considered not interesting and were eliminated, resulting in reduction of about 50% of transaction data.

Sampling of data with reference to representative cases of customers in terms of area, activity and class (insolvent or solvent) was performed. This resulted in a data set concerning the 2% of transactions and customers of the company.

4.4 Feature and Algorithm Selection for Data Mining

At this phase the data mining algorithms to be used are defined (in our case decision trees, neural networks and discriminant analysis) and the transformed dataset of the previous phase is further reduced by selecting the most useful features in adequate form for the selected algorithm. This feature selection is based mostly on automatic techniques, however domain knowledge is used for interpretation of the selected feature set. Also this process is used for verification of the previous phase assumptions, so if certain features do not prove to be discriminating factors then new attributes should be deduced and tested. It should also be mentioned that this feature selection process is often interleaved with the data mining process, since many algorithms select the most relevant features during the training process. In our case a stepwise discriminant analysis was used for feature selection.

4.5 Data Mining

Training a classifier using the cases of the collected data is considered the most important phase of the process. Depending on the mining algorithm selected, the de-rived knowledge can be interpreted by domain experts. For instance the rules defined by a decision tree can be inspected by domain experts. Also the weights related to the input

variables of a neural network reflect their relevant importance in a specific net-work. This is related to the performance of the model.
Extensive experiments often take place using a trial and error approach, in which the contribution of the classes in the training dataset and the input features, as well as the parameters of the data mining algorithm used, can vary. The performance of the deduced models indicate which of the models are most suitable for the knowledge-based system.

In an extensive experimentation that took place in the frame of our case study, 62 features were included in the original data set. Subsequently, 5 different datasets where created that where characterised by different distribution of the classes (S)olvent/ (I)nsolvent customer. These distributions were the following: (I/S: 1:1, 1:10, 1:25, 1:50, 1:100)

A ten-fold validation of each data mining experiment was performed, by redistributing the training/testing cases in the corresponding data sets. This way 50 classifying decision trees were obtained. By inspecting the features that have been used in these experiments, we selected the 20 most prominent, shown in Table 1.

Table 1. Most popular features used in the 50 classifiers

Feature	***Feature description***	***n.***
NewCust	Identification of a new connection	50
Latency	Count of late payments	50
Count_X_charges	Count of bills with extra charges	50
CountResiduals	Count of times the bill was not paid in full	50
StdDif	Std Dev. of different numbers called	50
TrendDif11	Discrepancy from the mov. avg. of four previous periods of the count of different numbers called, measured on the 11th period.	50
TrendDif10	Idem for the 10th period	50
TrendDif7	Idem for the 7th period	50
TrendDif6	Idem for the 6th period	50
TrendDif3	Idem for the 3rd period	50
TrendUnitsMax	Maximum discrepancy from the moving average in units charged over the fifteen 2-week periods.	45
TrendDif5	Idem for the 5th period	43
TrendDif8	Idem for the 8th period	40
Average_Dif	Average # of different numbers called over the fifteen 2-weeks period.	39
Type	Type of account, e.g. business, domestic etc.	33
MaxSec	Maximum duration of the calls in any 2-week period during the study period.	31
TrendUnits5	Discrepancy from the moving average of the units charged, measured on the 5th period.	28
AverageUnits	Average # of units charged over the fifteen 2-weeks periods.	23
TrendCount5	Discrepancy from the moving average of the total # of calls, over the fifteen 2-week periods.	21
CountInstallments	Count of times the customer requested payment by instalments.	18

In Table 1, one may observe that the time-dependent feature most frequently used was the one related with the dispersion of the telephone numbers called (*TrendDif*, *StdDif* etc., 9 occurrences). This is a derived feature, proposed by the domain experts as discussed above, that could not possibly be defined without the domain experts

Case distribution 1:1

If (StdDif<0.382952541) And (MaxSec<1086)
Then
INSOLVENT (confidence 1.4%)

If (StdDif<0.382952541) And (MaxSec>1086) And (ExtraDebt>=1.5)
Then
INSOLVENT (confidence 100%)

If (StdDif>=0.382952541) And (TrendCountMax>=4.625)
Then
INSOLVENT (confidence 5.36%)

Case distribution 1:10

If (CountXCharges<1.5) And (NewCust<0.5) And (TrendDif11<-0.625) And (TrendSec3<-1863.75)
Then
INSOLVENT (confidence 0%)

If (CountXCharges<1.5) And (NewCust>=0.5) And (CountResiduals>=0.5) And (TrendDif7<-0.625)
Then
INSOLVENT (confidence 12.5%)

If (CountXCharges<1.5) And (NewCust>=0.5) And (CountResiduals>=0.5) And (TrendDif7>=-0.625) And (StdDif<0.487950027) Then
INSOLVENT (confidence 10.93%)

If (CountXCharges>=1.5) And (StdDif>=0.305032313) And (TrendUnitsMax>=121.25) And (TrendUnits6<-2.375) And (TrendDif10<-0.125)
Then
INSOLVENT (confidence 12.26%)

If (CountXCharges>=1.5) And (StdDif>=0.305032313) And (TrendUnitsMax>=121.25) And (TrendUnits6>=-2.375) then
INSOLVENT (confidence 7.6%)

Case distribution 1:25

if (CountXCharges<2.5) AND (NewCust>=0.5) AND (CountResiduals>=0.5) AND (TrendCount5>=-1.25) AND (TrendDif6<-0.375)
then
INSOLVENT (confidence 25.8%)

if (CountXCharges<2.5) AND (NewCust>=0.5) AND (CountResiduals>=0.5) AND (TrendCount5>=-1.25) AND (TrendDif6>=-0.375) AND (TrendCount5<1.375) AND (Type<55.5)
Then
INSOLVENT (confidence 55.03%)

if (CountXCharges>=2.5) AND (TrendDif3<-0.125) AND (TrendUnitsMax>=222.625) Then
INSOLVENT (confidence 9.49%)

Fig. 1. Knowledge in form of rules, determining *Customer Insolvency*

contribution. This table demonstrates the important role of the domain experts in suggesting meaningful features during this phase.

4.6 Evaluation and Interpretation of Learned Knowledge

Evaluation of the learned knowledge usually involves measuring the performance using a test data set. However this also involves knowledge interpretation, as discussed in the previous section, which involves domain experts. Knowledge interpretation can be based on the performance on test cases and on inspection of the derived knowledge if adequate knowledge representation formalism has been used. The evaluation criteria

for the learned knowledge performance may be related to business objectives as defined by domain experts. An example of evaluation criteria is described in this section.

In figure 1 the knowledge in the form of rules, classifying the minority class cases (INSOLVENT customers) are exposed. It may be noticed that there is a considerable deviation in the parameters contributing to each of the rules, while the measure of performance of the rules vary considerably as indicated by the confidence measure expressing the rule performance in the test data set.

The criteria used for quantitative evaluation of learned knowledge in our case, as suggested by the domain experts, were different than the usual overall success rate and the specific class success rate indices usually applied in this kind of experiments. The domain experts suggested the following two criteria in our case study:

- The precision of the classifier, which is defined as the percentage of the actually insolvent customers in those, predicted as insolvent by the classifier.
- The accuracy of the classifier, which is defined as the percentage of the correctly predicted insolvent out of the total cases of insolvent customers in the data set.

These measures in problems of imbalanced class distributions, like in our case, in which the incidents of insolvent customers are very rare compared to those of solvent ones, seem more appropriate for measuring the effectiveness of the induced knowledge. By introducing these criteria, we discovered that the learned knowledge, despite of the fact that had very high success rates both overall and in specific classes, it did not meet the business objectives as these were defined by the Telecommunication Company (i.e. the requested measure of success was precision > 80% and accuracy > 50%).

An example of such a classifier is presented in the following table 2. In this table the performance of the classifier is shown in the testing data set. From this table one can see that the performance of this particular classifier is over 90 % in the majority class and over 83% in the minority class. However the precision is 113/2844= 3.9% and the accuracy is 113/136= 83%, thus making the performance in terms of the business objective set, not acceptable.

Table 2. Performance of classifier C1-3 for the insolvency prediction problem

	Predicted cases		
	Category	**Insolvent (0)**	**Solvent (1)**
Actual cases	**Insolvent (0)**	113 (83.1 %)	23 (16.9%)
	Solvent (1)	2731 (9.8 %)	25081 (90.2 %)

4.7 Fielding the Knowledge Base

This stage is essential in knowledge-based system development project, while this is often omitted in data mining projects as considered outside the scope of the data mining experiment. During this phase the learned knowledge is combined with other domain knowledge in order to become part of an operational decision support system, used by the company that commissioned the KDD project. The domain knowledge plays an important role during this stage. Usually the learned knowledge is just a part of this knowledge-based system, while heuristics or other forms of knowledge are often used as pre- or post-processors of the learned knowledge. In our case, the domain experts have suggested that the customers classified as insolvent, should be examined in more detail in terms of the amount due, the percentage of this amount that is due to third telecommunication operators, previous history of the customer etc, attributes that did not participate in the classification algorithm's decision, yet important for taking measures against the suspected insolvency.

In the fielded knowledge based system important aspects are also the available means for convincing the decision-maker for the provided advice. This can be achieved by providing explanation on the proposed suggestion or visualizing the data and the knowledge used, as suggested by many researchers, see Ankerst et al [12], Brachman & Anand [2], etc.

5 Conclusion

This paper has focused on the role of domain knowledge in a data-mining project. Eight distinct phases have been identified in the process and the role of the domain experts in each one of them has been discussed. In summary this role is shown in Table 3.

From this table and the discussion of the previous section it can be seen that while it is true that the domain knowledge plays a crucial role mostly in the initial and final stages of the process, it has contributed to some degree in all the phases of the project. If one takes also in consideration that the data mining phase (e), that necessitates comparatively little use of domain knowledge, accounts usually for not more than the 5% of the effort of such project, according to Williams and Huang [13], it is the most demanding stages of the process in which the domain experts and the domain knowledge participate mostly.

A conclusion of this study is that the data mining projects cannot possibly lead to successful knowledge-based systems, if attention is not paid to all the stages of the process. Since the domain knowledge plays such a crucial role in most of these stages, one should consider a data-mining project as a knowledge-driven process.

More support and adequate tools are therefore needed to be devised, which model the domain knowledge and track the contribution of domain experts that influence the assumptions made and the decision taken during the process.

Table 3. Overview of use of domain knowledge in a data mining project

stage	*Use of Domain Knowledge (DK)*	*Type of DK*	*Tools used*
(a) Problem definition	HIGH	Business and domain knowledge, requirements Implicit, tacit knowledge	
(b) Creating target data set	MEDIUM	Attribute relations, semantics of corporate DB	Data warehouse
(c) Data prepossessing and transformation	HIGH	Tacit and implicit knowledge for inferences	Database tools, statistical analysis
(d) Feature and algorithm selection	MEDIUM	Interpretation of the selected features	Statistical analysis
(e) Data Mining	LOW	Inspection of discovered knowledge	Data mining tools
(f) Evaluation of learned knowledge	MEDIUM	Definition of criteria related to business objectives	Data mining tools
(g) Fielding the knowledge base	HIGH	Supplementary domain knowledge necessary for implementing the system	Knowledge-based system shells and development tools

A concluding remark is that in terms of the actors involved in the process, next to the experts related to data analysis, data mining, data warehousing and data processing in a prominent position there should be put the domain experts that should participate actively and guide the process.

Acknowledgements. The research reported here has been funded under project YPER97-109 of the Greek Secretariat of Research and Technology. Special thanks are also due to the group of scientists of OTE S.A. that supplied us with data and continuous support. Special thanks to IBM for providing licenses of the DB2© and Intelligent Miner© products under their academic support programme. Also special thanks are due to the constructive comments of the anonymous reviewers on earlier draft of this paper.

References

1. Langley P., Simon H.A., Applications of Machine Learning and Rule Induction, Com. of the ACM, 38 (11), (1995), 55-64.
2. Brachman R. Anand T., “The Process of Knowledge Discovery in Databases: A Human-Centered Approach”, Advances in Knowledge Discovery & Data Mining, AAAI Press & The MIT Press: California, 996, (1996), 37-57.
3. Domingos P., “The Role of Occam's Razor in Knowledge Discovery”, Data Mining and Knowledge Discovery, an International Journal, Kluwer Academic Publishers, Vol.3, (1999), 409-425.

4. Yoon S.-C., Henschen L. J., Park E. K., Makki S., Using Domain Knowledge in Knowledge Discovery, Proc. ACM Conf. CIKM '99 1 l/99 Kansas City, MO, USA, pp. 243-250.
5. Anand S. S., Bell D. A., Hughes J. G., The Role of Domain Knowledge in Data Mining, Proc. ACM CIKM '95, Baltimore MD USA, pp. 37-43.
6. Van Heijst G., Schreiber G., CUE: Ontology Based Knowledge Acquisition, Proc. 8th European Knowledge Acquisition Workshop, EKAW 94, vol 867 of Lecture Notes in AI, pp. 178-199, Springer-Verlag, Berlin/Heidelberg (1994).
7. Wielinga B.J., Schreiber A.T., Breuker J.A., KADS: A modelling approach to knowledge Engineering, Knowledge Acquisition, 4(1), 5-53 (1992).
8. Fayyad U.M., Piatetsky-Shapiro G., and Smyth P., The KDD Process for Extracting Useful Knowledge from Volumes of Data, Communications of the ACM, 39(11), (1996)
9. Liebowitz J., Knowledge management and its link to artificial intelligence, Expert Systems with Applications 20, (2001) 1-6
10. Gur Ali, O.F., Wallace, W.A., Bridging the gap between business objectives and parameters of data mining algorithms, Decision Support Systems, 21, (1997) 3-15
11. Daskalaki S., Kopanas I., Goudara M., Avouris N., Data Mining for Decision Support on Customer Insolvency in Telecommunications Business, European Journal of Operations Research, submitted (2001)
12. Ankerst M., Ester M., Kriegel H-P, Towards and Effective Cooperation of the Computer and the user in Classification, ACM SIGKDD Int. Conf. on Knowledge Discovery & Data Mining (KDD'2000), Boston, MA (2000)
13. Williams G. J. and Huang Z, Modelling the KDD Process, A Four Stage Process and Four Element Model, TR DM 96013, CSIRO, Canberra, Australia (1996)