A TOOL FOR ECONOMIC PLANS ANALYSIS BASED ON EXPERT KNOWLEDGE AND DATA MINING TECHNIQUES

M. Baglioni
Department of Computer Science, University of Pisa
Largo B. Pontecorvo, 3 – 56125 Pisa Italy

B. Furletti
Department of Computer Science, University of Pisa
Largo B. Pontecorvo, 3 – 56125 Pisa Italy

F. Turini
Department of Computer Science, University of Pisa
Largo B. Pontecorvo, 3 – 56125 Pisa Italy

ABSTRACT

The project introduces an application for the analysis of business documents, based on data mining techniques and expert knowledge. The application is an integrated tool that combines economic theory and computer science technology in order to evaluate business documents (BDs) of start-up companies in order to foresee the success/failure of innovative projects. The Expert knowledge, represented by means of Bayesian Causal Maps, is used to drive and improve the classification process that is the core of the prediction strategy provided (Baglioni et al., 2005). In case of a negative result (failure), the system is also able to investigate the causes and to give, if possible, a suggestion for improving the plan.

KEYWORDS

Knowledge Discovery, Classification, Bayesian Causal Maps.

1. INTRODUCTION

After the introduction of stricter laws and regulations in the field of rating systems for banks and financial institutes, the design and construction of automatic control tools supporting decision makers is becoming more and more necessary. Also businessmen will need to make a sort of self-assessment to understand and deeply analyze their financial and economic position, the ability of the company to realize the projects, and the actual possibility of obtaining credit.

The tool we are proposing is part of a set of automatic instruments able to help businessmen, venture capitalists and financial institutes in the evaluation phase. In particular, the system addresses Business Documents (BDs) classification of start-up companies, which plan an innovative project. The aim of the system is the evaluation of the ability of the company in realizing the project on the basis of the submitted plan, and therefore the prediction of the success/failure of the project.

This kind of information is extremely important both for supporting the decision of granting a credit and for supporting self-assessment. By self-assessment the businessman can control the adequacy of the plan and, in case of negative prevision, can try to rearrange the proposal.

In this context, a solid economic model is important in order to support the automatic tools. To this purpose we use economic rules extracted from an economic data model coded as a Bayesian Causal Map (BCM) (Kemmerer et al., 2002). Summing up, our system uses three different reasoning techniques: the deductive technique provides the possibility of exploiting general evaluation rules derived from economical theories; the inductive technique, mainly based on statistical learning, provides learning and adaptation capabilities, and the abductive technique provides explanation capabilities.

The classification of business documents, and business plans in particular, will be tackled by the knowledge extraction and management techniques provided by the system. The idea is to exploit general economic knowledge as the deductive component, and inductive knowledge discovered from data mining analysis in order to classify and foresee the possibility of success of the project.

The predictive capability, along with the ability of finding and explaining the weak points of the plan under analysis can provide a robust and profitable tool for self-assessment.

The next section provides a quick overview of the technical background. Sec. 3 discusses the kind of economic knowledge embedded in the system, while sec. 4 presents the architecture and its main components. Sec. 5 deals with the crucial point of explaining why a plan is evaluated as doomed to failure. Then sec. 6 presents some experimental results and sec. 7 derives the conclusions and outlines the future of our research.

2. TECHNICAL BACKGROUND

One of the main characteristics of this tool is its ability to combine background (expert) knowledge and collected data in the construction of a classification model. In particular, the background knowledge is represented by means of BCMs. A BCM is a directed graph that connects concepts via a cause-effect relationship. In this kind of graph, arcs connecting related concepts are associated to a probability measure of the strength with which these concepts are related. BCMs are obtained by merging causal maps (Eden et al., 1992; Kemmerer et al., 2002), which are used to represent human believes, and Bayesian Networks (Mitchell, 1997). BCMs have two interesting properties: the first one states that concepts (nodes) connected by an arc are dependent (Causal Maps), and the second one states that concepts (nodes) that are not connected are conditionally independent (Bayesian Networks). This combination allows us to represent both if a concept influences another one and the strength of the influence.

Our classification model (Baglioni et al., 2005) is obtained by modifying the algorithm C4.5 (Quinlan, 1993). In a few words, the *Classification* is the process of finding a model that describes and partitions data classes or concepts. The model will be used to predict the class of objects whose class label is unknown. The model we chose is a decision tree, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and the leaves represent the class distribution. The C4.5 algorithm chooses the best node (attribute) among all the attributes at its disposal as the tree root. Once the root node is known, the algorithm repeats the same process on a subset of the values that is related to a specific value (set of values) of the attribute previously chosen. It is in this step that we can force the algorithm to assume that the values of a specific attribute are distributed according to the expert knowledge (if any), rather than according to the distribution implied by the data; it is in this way that the computation of the best next node is affected by the domain knowledge. The domain information is given by means of domain rules extracted from a BCM; this rules are associated to a probability value derived from the probability measures stored in the BCM. For a deeper treatment, please refer to (Baglioni et al., 2005).

3. ECONOMIC DATA MODEL

Which is the interesting information to extract from BDs has been defined by experts in the economic field. They pointed out six macro categories within which the items (attributes) can be partitioned:

1. *Company managers and management systems*: it contains information about the company managers and about the company organizational systems;
2. *Commercial dimension*: it contains information about the reference markets, the relationship between the company and the customers and the know-how of the person in charge of sales;

3. *Technical-production dimension*: it contains information on production agreements, patents, employer skills, skills of the person in charge of production and products design, and the degree of obsolescence of the equipments;
4. *Relationship with the Environment*: it contains information about the geographical concentration and the infrastructures of the company;
5. *Competitors analysis*: it contains information about providers, customers, potential customers, and competitors;
6. *Economical-financial analysis*: it contains information about the financial and economic structure, and the income of the company.

As mentioned above, the experts also provided us a with set of item correlations that we coded as a BCM. This net is an useful means to represent the overall influence and their weights.

4. THE SYSTEM

As mentioned before, our system offers a set of tools to support the evaluation of BDs. These instruments are based on both the analysis of past similar cases and expert knowledge. To realize this framework, we organized the system in 5 main components as it is shown in Figure 1.

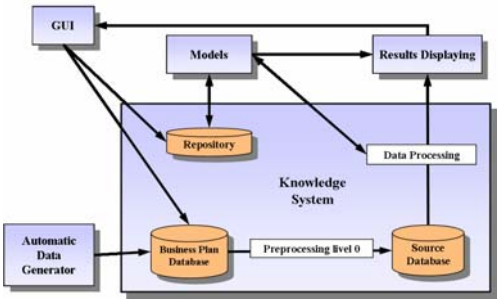

Figure 1. The System Architecture.

- The **GUI** is the Graphical User Interface, by which the user can insert data (historical data and new BDs, new economic models), make queries and get all kinds of system outcomes.
- **Models** is the component for building data mining and business models.
- The **Knowledge System** contains the data repository, which contains economic and data mining models, and the components for the phase of data pre-processing and processing. The *Business Plan DataBase* component contains all original instances of BDs, while the *Source DataBase* component contains the instances of BDs after pre-processing. The *Preprocessing level 0* module performs filtering and aggregation of data, whilst the *Data Processing* model handles the main computation. In this last phase the system checks the consistency of values and classifies the new BD. In case of negative result, it tries to suggest alternatives.
- **Results Displaying** is the component for translating the output into a form readable by the user.
- **Automatic Data Generator** is a temporary component we used to fill up the knowledge base in the absence of real data.

4.1 Automatic Data Generator

The economy experts have defined an economic model in order to generate values as similar as possible to the empirical ones. The probability distributions we consider for the item generation can be histograms, uniform distributions, gaussian distributions, or it can be given by means of relations pointed out by the experts. If the relation has a single item in both the left hand side and the right hand side, we consider the interval [0,..,1] and we split it in two sub-intervals: one from 0 to the probability expressed in the relation, and the other containing the remaining values. Then we generate a random real number between 0 and 1, and the value of the generated item is determined by the subset in which the generated item lies. We do the same

when the left hand side of the relation has more than one item: we consider the average of the probabilities and we proceed as before. When, instead, we consider a relation with more than one item in the right hand side, we have to proceed as it follows: we generate a natural number between 1 and the number of items in the right hand side of the relation to determine which will be the item to generate a value for. Then we generate the chosen item as it were a single item. If the generated value is different from zero, we set the values for the other items to zero, and we finish the generation for this relation. If it is zero, we start again the process until we have an item with value different from zero.

For example consider the item generation process of the first two groups of items within category one. This category is composed of five groups of items. The first group contains only the element corresponding to the number of the company managers. The value for this item depends also on the kind of company we are generating a value for. Hence we first generate the kind of company (small and inside the reference market, small and outside the reference market, medium and large) and then a number from 0 to 1 to discriminate the item value. The second group is composed of six items related to the age of all the managers in the company. The age of each manager is generated by exploiting the distribution $N(43, 20)$, and then the right value for the items belonging to the second group is chosen. For the remaining subgroups of this category and the other categories the generation process is quite similar.

Once the data have been generated, they are prepared for extracting the classifier. The kind of pre-processing we applied is substantially aggregation, value consistency verification, and determination of the value (using economical criteria based on item values and heuristics) for the target attribute. For a deeper treatment, please refer to (Baglioni, 2005) and (Fornasari, 2003).

On the pre-processed attributes, we are given a map of relations built by an expert in the economic field. This map describes the relationships that hold, independently of the data in the considered context. Each item occurring in the map refers to the category it belongs to, and to the group within the category. Because of space limitations we are not going to show neither a picture of the map, nor a list of the items it is composed of. More information about the BCM can be found in (Baglioni, 2005).

4.2 Data Processing

The data processing is the phase in which a new instance of BD is classified to obtain a prediction of success/failure. In case of a negative result, the system is able search for the causes, and to suggest a way to modify the plan.

5. INQUIRING ABOUT FAILURE

An interesting topic of research consists in trying to help businessmen in modifying their BDs whenever the system classifies them as “no success”. The aim is to provide a general idea of the causes for the negative response. However, it is not granted that the modified BD will be classified as successful by the system. Variables of a BD are strongly related and the quality of our explanation process depends heavily on the economic model we are given. So far, we got only some simple constraints and high level rules, that describe what type of variables can be modified (and how to modify them). For example, variables describing the external environment, the market and so on, are invariant and do not depend on the management company. Thus these values cannot be rearranged. Variables describing production agreements, innovation degree of the company etc. can instead be modified.

To solve this problem, we first analyze the path (root-to-leaf), traversed by the BD during the classification step. The analysis follows a bottom-up approach and starts checking the variable and its value from the leaf up to the root. If the value of the variable leads to a negative response, we try to suggest a range of alternative values according to rules and constraints given by the expert. If no alternatives are available, the analysis is recursively applied one step up to the root. The process terminates when at least one alternative value is found. No suggestions are given if we reach the root and if there is not a “good” alternative value for the variable in this position. Figure 2 shows the steps of the analysis for the following case.

Example 1: Let us consider the classification tree in Figure 2 (first tree), and suppose we have to classify the instance *Inst* composed by the following Attribute-Value pairs separated by “;”:

Inst: $A \leq a_1$; $B=b_1$; $C=c_1$; $D>d_1$; $E=e_2$; $F=f_1$; $G \leq g$; …

The dotted line of Figure 2 (first tree) leads to the leaf labeled *fail* that stands for a negative prediction.

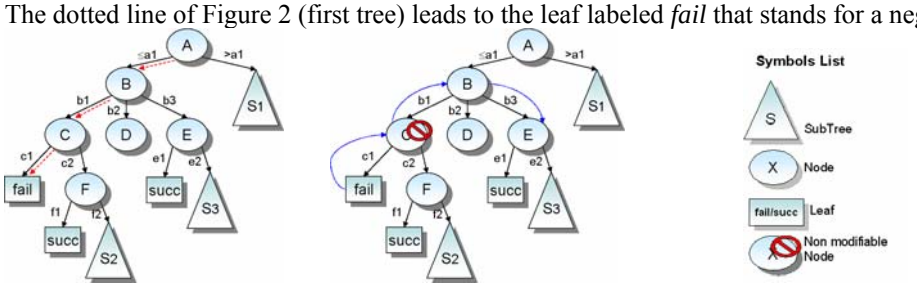

Figure 2. Example of the classification path (first tree) and the steps of the bottom up analysis (second tree).

The inquiring process, shown in Figure 2 (tree on the right), starts from the leaf parent node *C*. Suppose that *C* is a "non modifiable" attribute, then the process continues by analyzing the node *B*. Suppose that *B* is a "modifiable" attribute and that the constraints, given by the experts, impose the value b_3 as the available one. Since from attribute *E* the tree proceeds with a successful leaf ,and a sub-tree *(S_3)*, the process can finish. The suggestion is to change the value of attribute *B* (if possible) with b_3. Modifying a value of an attribute can lead to a plan re-arrangement and then a new classification process will be necessary.

6. SOME RESULTS

Figure 3 shows the Analysis Panel that allows the user to create a new model (that is a new classification tree), to load an already created model, and to proceed with the analysis of the instance. In this case, the picture shows the final phase of the analysis in which the system suggests a modification after classifying the project as not successful.

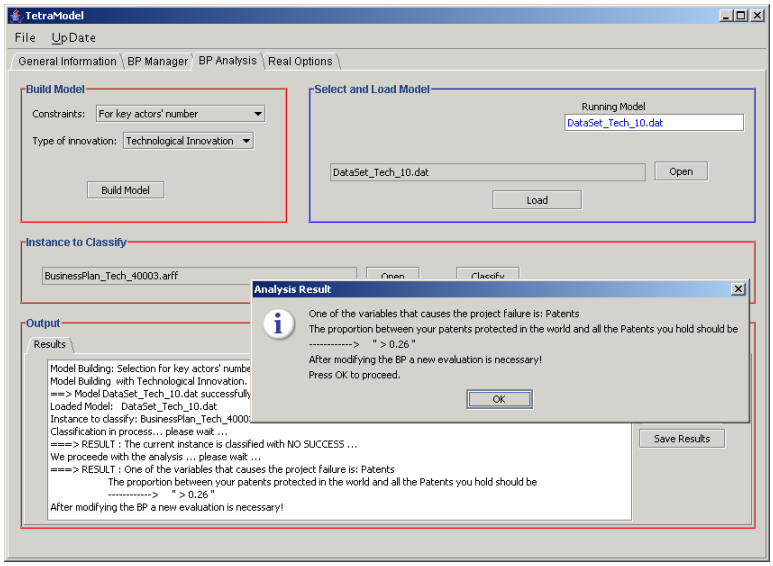

Figure 3. Graphical Interface: Analysis Panel.

The experiments have been carried on according to the following steps: *synthetic generation* of about 40000 BDs according to the techniques described in sec. 4.1; *elicitation of expert knowledge* in the form of BCMs as described in sec. 2; *generation of 3 classification models* by using our algorithm; *classification* of eleven real BDs by using the three classification models according to majority voting.

The classification accuracy and the classification results are presented in Figure 4 (a) and (b). Figure 4 (c) shows a comparison of the results of our tool w.r.t. pure C4.5 when considering each classifier separately. Although the results are not very accurate, we can see that domain knowledge helps in improving them w.r.t.

plain induction, as performed by C4.5. Both better data (which can be obtained by harvesting more real BDs and using them in place of or, even better, along with the synthetic data) and refining the expert knowledge coded in the BCM, can be helpful for improving them. This is an ongoing project and we are actively pursuing both alternatives.

7. CONCLUSION

Supporting decision making in planning the development of companies is a very hard task. On the other hand, risk assessment and self-assessment are essential for companies, especially small/medium enterprises, that need to ask for financial credit, that plan to develop new lines of business, that plan to internationalize their business.

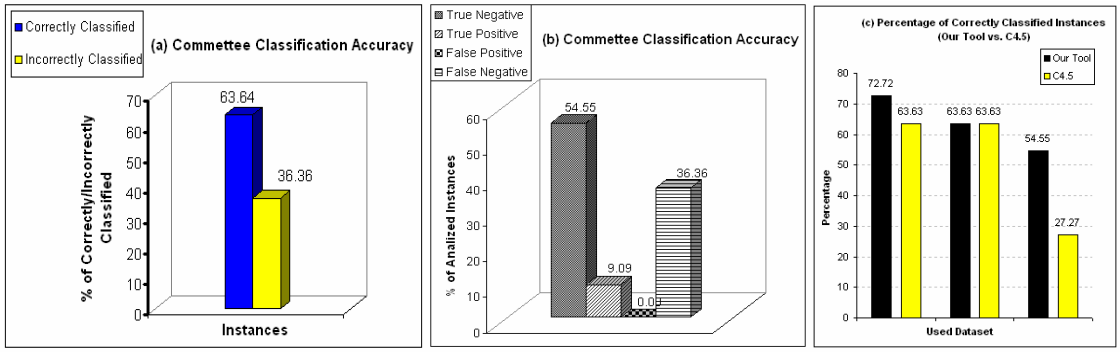

Figure 4. Classification Analysis.

In addressing this kind of very complex tasks, first in a national funded project, named TetraModel, and now in two new IP projects within the sixth framework of the European Union, we realized that traditional inductive methods alone could not provide a satisfactory solution. We found the absolute need for eliciting expert knowledge in three respects: first to construct a sensible model of the data to be extracted from business documents, second to integrate the induction process when extracting classification models with heuristic rules taken from experts, third for driving the search for possible adjustments for failing plans.

This is an ongoing work, and we cannot claim completely successful results, so far. However, the results we have got so far make us confident that the choice of integrating expert knowledge and the inductive approach can be a good way to solve the complex and difficult problems occurring in the field of economic assessment.

REFERENCES

Baglioni, M. et al, 2005. DrC4.5: Improving C4.5 by means of prior knowledge. *Proceedings of the 2005 ACM Symposium on applied computing.* Santa Fe, New Mexico, pp. 474-481.

Baglioni, M., 2005. *Representation and exploitation of Domain Knowledge in KDD Environment*, PHD thesis. Computer Science Department – University of Pisa, Italy.

Eden C. et al, 1992. The analysis of cause maps*. Journal of Management Studies,* 29(3), pp. 309-323.

Fornasari, F., 2003. *Progettazione di metodi per il pre-processing dei dati*. Deliverable of the TetraModel project (available only in Italian).

Han, J. and Kamber, M., 2001. *Data Mining: Concepts and Techniques.* Morgan Kaufmann Publishers, San Francisco, USA.

Kemmerer B. et al, 2002. Bayesian Causal Maps as Decision Aids in Venture Capital Decision Making: Methods and Applications. *In proceedings of the Academy of Management Conference.*

Mitchell, T., 1997. *Machine Learning.* McGraw Hill Publisher.

Quinlan, J. Ross, 1993. *C4.5: programs for machine learning.* Morgan Kaufmann Publishers Inc.