BUSINESS BANKRUPTCY PREDICTION BASED ON HYBRID CBR MODEL

Bruno Trstenjak  
Department of Computer Science  
Polytechnic of Medimurje in  
Cakovec  
Bana J.J. 22b, Cakovec,  
Croatia  
Email: btrstenjak@mev.hr

Dzenana Donko  
Department of Computer Science  
Faculty of Electrical Engineering  
Zmaja od bosne bb, Sarajevo  
Bosnia and Herzegovina  
Email: ddonko@etf.unsa.ba

KEYWORDS  
Bankruptcy Prediction, Hybrid CBR model, Case-based reasoning, Domain driven model.

ABSTRACT

Predicting bankruptcy and early identification of the financial crisis nowadays has become a field of particular interest in which various studies have been conducted. The consequences of business bankruptcy have a negative impact on the whole society. Signs of financial distress can be detected much earlier before bankruptcy occurs. For this reason, a variety of scientific methods have been developed for timely detection of the difficulties in the business.  
The main purpose of this paper is to analyze quality prediction of business bankruptcy using the novel hybrid Case-based reasoning (CBR) model with a new approach of data classification. The hybrid model in its algorithm has integrated several methods of machine learning: Information Gain, K-means and Case-based reasoning. The principle of work of the hybrid model is based on the experts knowledge base, library of past cases and determined weight values of object features. The model in the process of prediction uses the entire information as a basic value and does not use features reduction methods. jCOLIBRI is used as a framework for the development of our hybrid CBR model and algorithm. Our experiment has achieved promising results and provides a good basis for future research.

INTRODUCTION

Today in the modern world business bankruptcy has become commonplace. Business bankruptcies and their frequency always have a significant negative impact on the whole society, especially on the social and economic component. The causes of business bankruptcy have been discussed by many authors, trying to identify indicators that will signal the possibility of negative scenarios in the business. Paul J. FitzPatrick identified five stages leading to business problems: incubation, financial embarrassment, financial insolvency, total insolvency, and confirmed insolvency (Fitzpatrick P.J. 1932). All these and similar efforts are focused at defining features that describe the business object and on the basis of their values predict business bankruptcy. In order to prevent and predict such negative events different kind of research has been conducted. The machine learning is one of the scientific fields in which the concept of prediction has been explored and various algorithms have been developed in order to archive better prediction accuracy (Baldi et al. 2000), (Pandya and Virparia 2013). Modern bankruptcy prediction models often use various statistical analysis and various methods of machine learning in the field of data classification. In the studies based on statistical analyses, the methods such as logical regression, multivariate discriminant analysis and factor analysis were used (Ohlson 1980), (West 1985). Somewhat different approach in prediction are provided by the use of machine learning methods such as neural networks (Ravisankar et al. 2010), support vector machines (Jong and Kwon 2010), Bayesian network models (Sun and Shenoy 2007) and many others. To achieve the highest prediction accuracy, models that use several machine learning techniques fused in a single algorithm were developed. Such models are called hybrid models, and their complex algorithms regularly achieve better classification results than individual methods (Ping and Yongheng 2011), (De Andrés et al. 2011). In order to achieve the highest possible prediction accuracy we have developed newer hybrid model and algorithm based on Case-based reasoning (CBR) method of classification. Our hybrid model provides a slightly different classification approach. The model has embedded three distinct processes with associated algorithms: generating weight values of features, clustering and classification. Information Gain (IG) is a filter selection algorithm that is used for ranking the input attributes. The large number of features that are included in the prediction model can influence on the speed of prediction. The irrelevant features can also have influence on the prediction accuracy. IG measures entropy of features and ranks them according to the measured values (Beniwal and Arora 2012). For the purposes of classification process K-means algorithm is used (An et al. 2008). K-means is one of the simplest unsupervised learning algorithms used for solving clustering problems. Clustering is the process of dividing data into smaller groups, clusters, associated on the basis of common properties. K-means algorithm is used to optimize the
clustering data, prepare for classification phase, all in order to improve the classification accuracy.

The third process in the hybrid model is classification of objects. For the purposes of classification CBR method was used. CBR method in the classification process uses the methodology of solving new problems by determining the similarity from the past. Experiences are essential for CBR method and principle of its work. All collected experiences are written in the form of cases. CBR investigates cases from the past, and on the basis of similarity, functions propose solution to the current situation (Hullermeier and Cheng 2013). The CBR method based on the four cycles proposed by Aamodt and Plaza [1994] is a commonly accepted process model.

The rest of the paper is organized as follows; Research objectives section briefly describes research objectives, Overview of hybrid CBR model section deals with the concept of a novel hybrid algorithm and the knowledge database. In the Case study section presents the achieved results of the case study and Conclusion section presents conclusions and directions for further development.

RESEARCH OBJECTIVES

The main goal of this study is to analyze quality prediction of business bankruptcy using the novel hybrid CBR model with a new approach of data classification and implement classification of objects based on the features weight values and database knowledge from the domain of problem. In the process of predicting, we attempt to preserve all the objects information without reducing the number of their features. We wanted to measure maximum prediction accuracy and explore the dependence between the accuracy of prediction and the size of the dataset. Detect weak and the good characteristics of the hybrid model in the process of prediction.

OVERVIEW OF HYBRID CBR MODEL

During the years of research in various domains of problems we have witnessed that experience and knowledge can play an important role in making the final decision. Sometimes decisions based on experience can give better results than decisions based on scientific mathematical methods and algorithms. All this led to the idea of combining the experience and knowledge with the machine learning methods, all with the goal to achieve high accuracy in prediction of business bankruptcy.

The algorithm of our hybrid model prediction is carried out in three internal processes: generating weight values of features, clustering and classification. The process of generating the weight values is performed during model initialization. A generator uses the IG method for ranking the features of objects and performs aggregation of results with the data from domain driven model. The hybrid algorithm with the aggregation of information forms a weight vector which forwards in the process of classification. Two first processes in the hybrid algorithm provide additional information and thereby strives to increase the accuracy of classification. Figure 1 shows the concept of a hybrid CBR model.

The weight generator

The main purpose of the generator is to determine the weight values of features, the features which describe the objects in the dataset. The generator is composed of two components: IG method for ranking object features and domain driven model which contains knowledge of experts from the problem domain.

![Figure 1: The structure of the hybrid CBR model](image)

The domain driven model presents the knowledge of experts and their opinions about the importance of features that describe the business object. In the knowledge base weight value for each feature is written. The value indicates the importance of each prediction feature based on the opinion of experts; strength impact of features on the final result of prediction. Various studies in the classification field showed that the features have different influence on prediction results (Janecek and Gansterer 2008). Weight values in domain driven model were determined by experts, based on their experiences and opinions. In our study, the values were determined by surveying experts from various fields of economy who work in our institution. The values are in the range from 0–1, Table 1 shows meanings of the values. The experiences which we have obtained during our research has shown that sometimes the method of ranking the importance of features and their weight values could be incorrect. Therefore, Domain driven model was included in the algorithm as a corrective factor in the process of defining
weight vector. The hybrid algorithm performs correction of weight values, which has computed using IG method.

Table 1: Values in Domain driven model

<table>
<thead>
<tr>
<th>Value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>The feature is not important for the process of prediction and is not used. The attribute is ignored.</td>
</tr>
<tr>
<td>0&lt; x &lt;1</td>
<td>The value represents the corrective factor in the process of generating a weight vector.</td>
</tr>
<tr>
<td>1</td>
<td>The feature has a maximum impact on the result of prediction. The feature must come into the process of prediction.</td>
</tr>
</tbody>
</table>

Entropy is a measure of disorderliness of the system. Information Gain (IG) method calculates the value of the features information. Value is defined as the amount of information, provided by the feature items for the class. IG uses the following expression for the calculation:

$$IG(Class, Feature) = H(Class) - H(Class | Feature)$$

where H is entropy, which is defined by using the following equation (1).

$$Entropy(S) = -\sum_{j=1}^{m} p_j \log_2 p_j ,$$

where p is probability, for which a particular value occurs in the sample space S.

Entropy value ranges from 0 to 1. Value 0 means that all variable instances have the same value, value 1 equals the number of instances of each value. Entropy shows how the attribute values are distributed and indicates "purity" of features. High entropy indicates the uniform distribution of attributes. Opposite to that, low entropy indicates the distribution which is concentrated around the point (Lei 2012).

Data clustering

K-means is one of the simplest unsupervised learning algorithms used for solving clustering problems. Let X={x_i, i=1...n} be the set of n dimensional objects which should be classified into k clusters, C={c_j, j=1...k}. The algorithm determines the quality of the clustering calculating square error between the mean of a cluster and the points in the cluster. The goal of algorithm is to minimize the sum of the squared error over all K clusters. The quality is determined by following the error function, as shown equation (2) (Jain 2010):

$$E = \sum_{j=1}^{k} \sum_{x \in C_j} |x_i - \mu_j|^2 ,$$

where E is sum of the squared error of all objects, \( \mu_j \) indicates the average of cluster C_j, \(|x_i-\mu_j|^2\) is a chosen distance measure between data point x_i and centroids value.

The algorithm can use different methods to calculate the distance (Euclidean, Manhattan, Minkowski, etc.) (Kouser and Sunita 2013).

Data classification

The principle of CBR method is based on solving new problems by observing the similarity with the previous solved problems. CBR method uses a problem-solving approach analogous to the way of problem solving by man when he draws on his experiences (Klein 1999). Each CBR system contains embedded library of past cases that have been resolved in the past. This is something like collecting life experiences in the domain of the problem. Each case represents a description of the problem with its associated solution. CBR method with built-in function of similarities tries to find the most similar case from the library. The retrieved cases from the library are used to suggest a solution. If the proposed solution is not satisfactory, method tries to revised selected cases and find a new solution. The method adds a new revised case to the cases library and thereby expands the knowledge base. The whole execution cycle of algorithm can be divided into four main steps (Richter and Weber 2013):

1. Retrieve - retrieval is the first step in the cycle. The algorithm tries to find the best matching case(s). The case that will be selected among the retrieved cases depends on the similarity function which is used.
2. Reuse - If the new problem situation is exactly like the previous one then algorithm reuses the old case. If the retrieved cases do not offer acceptable solution for a new problem, the algorithm performs the adaptation of retrieved cases.
3. Revise - This step starts when a solution is proposed to solve a new problem. The aim of revise is to assess the acceptability of proposed solutions, the newly formed case.
4. Retain - In the retain step there are new useful cases in the case base for future reuse. In this way the CBR system has learned a new experience and retains the knowledge gained from solving the new problem.

Which cases to be retrieved is decided based on a given similarity threshold. CBR performs measurement of similarity on the local and global level. Local similarity refers to the measurement of similarity between pairs of features. Global similarity refers to a comparison of the similarity between all the features that make up the object. Measuring similarity can be shown by the following equation (3) (Najib and Ahmad 2012):

$$Similarity(T,S) = \sum_{j=1}^{n} f(T_j, S_j) \times w_j ,$$

Where

T= target case
S= source case
n= number of features in each case
I= individual feature from 1 to n  
f= similarity function for features I in cases T and S  
w= importance weighting of feature I. 

CASE STUDY 

Dataset 

Business bankruptcy is a financial failure when a company or organization is unable to pay its financial obligations (Ohlson 1980). In order to implement the process of prediction it is necessary to detect the features by which the model will classify objects in the analysis. Features of objects which are used in prediction can be of financial and non-financial nature. In this research we used a validated and prepared datasets created by Jeffrey S. Simonoff (Simonoff 2003). Business companies have described with seven features. For the testing purposes we used a dataset of 50 instances. Table 2 shows an overview of the features (Simonoff 2003):

<table>
<thead>
<tr>
<th>No</th>
<th>Feature</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CN</td>
<td>Name of business entity.</td>
</tr>
<tr>
<td>2</td>
<td>WC/TA</td>
<td>Working capital as a percentage of total assets; the difference between current assets and liabilities, and is thus a measure of liquidity.</td>
</tr>
<tr>
<td>3</td>
<td>RE/TA</td>
<td>Retained earnings as a percentage of total assets; a measure of cumulative profitability over time, and is thus an indicator of profitability.</td>
</tr>
<tr>
<td>4</td>
<td>EBIT/TA</td>
<td>Earnings before interest and taxes as a percentage of total assets; a measure of the productivity of a firm's assets.</td>
</tr>
<tr>
<td>5</td>
<td>S/TA</td>
<td>Sales as a percentage of total assets; the standard capital turnover ratio, indicating the ability of a firm's assets to generate sales.</td>
</tr>
<tr>
<td>6</td>
<td>BVE/TL</td>
<td>Book value of equity divided by book value of total liabilities; ratio measures financial leverage, being the inverse of the debt to equity ratio</td>
</tr>
<tr>
<td>7</td>
<td>Class</td>
<td>B-Bankruptcy, NB-Non-Bankruptcy.</td>
</tr>
</tbody>
</table>

Table 2: Overview of features in the dataset 

Table 3:Overview of the aggregation results

<table>
<thead>
<tr>
<th>No</th>
<th>Feature</th>
<th>IG</th>
<th>DDM</th>
<th>Aggre.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>CN</td>
<td>1.00</td>
<td>0.0</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>WC/TA</td>
<td>1.18</td>
<td>0.73</td>
<td>0.96</td>
</tr>
<tr>
<td>3</td>
<td>RE/TA</td>
<td>0.55</td>
<td>0.60</td>
<td>0.58</td>
</tr>
<tr>
<td>4</td>
<td>EBIT/TA</td>
<td>0.45</td>
<td>0.67</td>
<td>0.61</td>
</tr>
<tr>
<td>5</td>
<td>S/TA</td>
<td>0.00</td>
<td>0.87</td>
<td>0.43</td>
</tr>
<tr>
<td>6</td>
<td>BVE/TL</td>
<td>0.36</td>
<td>1.00</td>
<td>0.68</td>
</tr>
</tbody>
</table>

Table 3 shows: IG - values calculated with IG method, DDM - values obtained from a questionnaire survey, AGG - results of aggregation and final weight values of features. Indicative situation can be observed in Table 3, the situation where the IG method marked the name of the business entity as an important feature in the dataset. In the domain driven model, that feature was marked as irrelevant for the prediction business bankruptcy. In the process of determining the final weight value of feature, the marked feature was excluded from the prediction process. The opposite situation was created for BVE/TL feature. The results of conducted survey indicate a BVE/TL feature as something that should be necessarily used in the process of prediction. For other features, the algorithm performs a correction by reducing or increasing their values. In order to evaluate contributions of the hybrid model, before testing the hybrid model an initial data classification has been performed. Classification of raw data was conducted using the Bayesian network method.

Table 4: Results of the initial classification

<table>
<thead>
<tr>
<th>Stratified cross-validation</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly Classified Instances</td>
<td>86 %</td>
</tr>
<tr>
<td>Incorrectly Classified Instances</td>
<td>14 %</td>
</tr>
<tr>
<td>Kappa statistic</td>
<td>0.72</td>
</tr>
<tr>
<td>Mean absolute error</td>
<td>0.14</td>
</tr>
<tr>
<td>Root mean squared error</td>
<td>0.3198</td>
</tr>
<tr>
<td>Relative absolute error</td>
<td>28.42 %</td>
</tr>
<tr>
<td>Root relative squared error</td>
<td>63.67 %</td>
</tr>
</tbody>
</table>

Generating a weight values 

In the initial phase of prediction, hybrid model begins with the generation of weight values for each feature from the dataset. Generating was performed in a few steps. IG method performs ranking of features and determines their importance, presence in the dataset. Parallel with this activity, algorithm collects information from domain driven model. In this study, the knowledge base is formed through a questionnaire conducted in the Polytechnic of Medimurje in Čakovec, Croatia. A survey was conducted over experts from various fields of economic science. Each expert assessed the importance of the features in the range from 0 to 5. Value 0 denotes that the feature is irrelevant for prediction and has no influence on business bankruptcy. Value 5 in the table indicates the attributes that are most important in the process of determining bankruptcy. On the basis of their opinions we formed a study knowledge database. The algorithm performs data aggregation by merging the results obtained by the IG method and the data from Domain driven model. Table 3 shows an overview of the aggregation results.

During the initial classification of raw data from a dataset it was determined that the feature "Company name" is an irrelevant feature. We performed classification with and without "Company name" feature. The obtained measurement results are identical in both cases. When logically considering the situation, a personal name can not
have an impact on the the outcome of prediction. Precisely because of such situations, a hybrid model uses knowledge and Domain driven model. To prevent incorrect classification due to an error in the definition of the input data. Table 4 shows obtained results of initial classification using Bayesian network method.

**Data clustering**

Parallel with the process of generating the weight values, hybrid algorithm performs clustering of objects using the K-means method. To achieve high accuracy of prediction it is necessary to define the optimal number of clusters. The initial number of clusters was determined using a simple rule of thumb shown in equation (4) (Madhusudhan 2012):

\[
k \approx \sqrt{\frac{n}{2}}
\]

where \( n \) is the number of objects.

According to equation (4) the initial number of clusters was determined and its value was 7. For measuring distances we used Manhattan function. K-means using a distance function measures dissimilarity between objects. Results of objects clustering are shown in Table 5:

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Clusters</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>11 (22%)</td>
</tr>
<tr>
<td>1</td>
<td>3 (6%)</td>
</tr>
<tr>
<td>2</td>
<td>10 (20%)</td>
</tr>
<tr>
<td>3</td>
<td>5 (10%)</td>
</tr>
<tr>
<td>4</td>
<td>7 (14%)</td>
</tr>
<tr>
<td>5</td>
<td>8 (16%)</td>
</tr>
<tr>
<td>6</td>
<td>6 (12%)</td>
</tr>
</tbody>
</table>

The results present the distribution of objects, i.e. a similarity between the business entities based on their features values. The results show that the objects are relatively uniformly represented in the dataset and dataset does not have extreme cases.

**Data classification**

Such prepared data were forwarded to the final stage of prediction, classification with the CBR method. For the purpose of the case study, we used jCOLIBRI framework (Recio-Garcia at el. 2008) adapted to the domain of the problem. In the jCOLIBRI framework we have integrated our own classes adapted for measurement prediction of bankruptcy. Classification results show that the initial value of the number of clusters was well determined. The classification with such prepared data shows very good final prediction results, presented in Table 6. The results from Table 6 show that better results are achieved using the hybrid model compared to the initial classification obtained by the Bayesian network method. The hybrid model has achieved extraordinary good classification results. The achieved improvement of prediction exceeded our initial expectations. The prediction accuracy over 99% is excellent.

<table>
<thead>
<tr>
<th>Stratified cross-validation</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Business Bankruptcy Accuracy: 99.6</td>
<td></td>
</tr>
<tr>
<td>Number of Cycles: 50</td>
<td></td>
</tr>
<tr>
<td>Time per Cycle: 15.26 ms</td>
<td></td>
</tr>
<tr>
<td>Total time: 763 ms</td>
<td></td>
</tr>
</tbody>
</table>

At the end of our study, we measured sensitivity of the hybrid model to the size of the dataset and number of cases. The original dataset consists of 50 objects. At the beginning of the testing and measurement accuracy of classification, we have reduced the number of objects to 40%. After that, the number gradually increased per step of 10 objects. Figure 2 shows obtained results of classification dependent on the number of objects.

![Classification accuracy](image)

The accuracy of prediction depending on the number of objects

From the achieved results it is evident that hybrid model achieves high accuracy of prediction also with a smaller number of objects. During gradual increase of the number of objects, the model has increased the prediction accuracy. The results present a relatively low sensitivity of the hybrid model to the dataset size. These model characteristics ensure its application in various fields of research.

**CONCLUSIONS**

In this paper we present a novel concept of hybrid model for prediction. The aim of our research was the development of a
new hybrid and Domain driven model based on CBR method. The model was focused on business bankruptcy prediction. According to the results achieved with a hybrid model, we can conclude that we have successfully designed a concept of a hybrid algorithm, combining the classical machine learning methods and experts knowledge from the problem domain.

The final results showed that with this model we can achieve significantly higher prediction accuracy compared to the conventional approach. The hybrid algorithm has proved stability in prediction. During testing, the model showed sensitivity in the prediction accuracy based on the defined weight values. Also, the model achieves high accuracy of prediction using a relatively small dataset. This feature of the model gives the possibility of its use in different areas which are not explored and where there is no large knowledge database. Our concept is certainly a good starting point for further development with various types of predictions. Future testing of different types of datasets will indicate the elements that will need to be modified, with the main purpose to achieve better results.

REFERENCES


BIOGRAPHIES

BRUNO TRSTENJAK is a lecturer at the Polytechnic of Medimurje in Cakovac, Croatia. In 2012, he completed MSc degree in the field of Computer sciences at the Faculty of Electrical Engineering in Sarajevo, BiH. Since 2008, he has worked as an academic member at the Medijumurje University of Applied Sciences, where he has been involved in lecturing several courses in fields of Computer Sciences. His area of interest covers: programming and application development, field of artificial intelligence, machine learning and data mining. He has been involved in the several EU projects in the field of education, as a project manager or a member. Recently, he is head of department at professional study program - Computer Engineering.

DZENANA DONKO is an associated professor at the Faculty of Electrical Engineering University of Sarajevo, BiH. She received her Ph.D. from the Faculty of Electrical Engineering, University of Sarajevo in 2004. Ms. Donko participated in many project and studies, some of them related to the UNDP development projects in BiH. Her current research interests include areas of object oriented analysis and data mining. She is author of many papers published at different journals and conferences. In several occasions she participated as a plenary speaker at the conferences presenting subject on data mining. She is also a member of many scientific and program committees on the different international conferences.