Intelligent System Design for Detect Bankrupt Corporations

Shirzad Mirdrikvandi, Amin Einipour
Department Of Accounting, Poldokhtar Branch, Islamic Azad University, Poldokhtar, Iran
sh.mirdrikvandi@gmail.com

Department Of Computer, Andimeshk Branch, Islamic Azad University, Andimeshk, Iran
a.einipour@gmail.com

Abstract
In recent times, every organization financial analysis provides the basis for understanding and evaluating the results of business operations and delivering how well a business is doing. This means that the organizations can control the operational activities primarily related to corporate finance. Bankruptcy prediction has drawn a lot of research interests in previous literature, and recent studies have shown that machine learning techniques achieved better performance than traditional statistical ones. One very common approach to bankruptcy prediction is by using classification techniques, based on key financial ratios. This approach also stands at the core of this paper, together with a GA. The purpose of combining the two is to find a suitable classification rule by which bankrupt firms can be separated from non-bankrupt ones. The results obtained by using the GA are benchmarked against the performance of Discriminate Analysis on the same data.

Keywords: Intelligent System, GA Algorithm, Corporate Bankruptcy.

1. Introduction
It might start with defaulting on an obligation, or when a company’s liabilities outweigh its assets. It’s called bankruptcy: a legally declared inability or impairment of ability of an individual or organization to pay their creditors. Its consequences are disastrous; no wonder that scientists and finance professionals have been trying, for over 50 years now, to develop efficient failure prediction models. One very common approach to bankruptcy prediction is by using classification techniques, based on key financial ratios. This approach also stands at the core of this paper, together with a GA. The purpose of combining the two is to find a suitable classification rule by which bankrupt firms can be separated from non-bankrupt ones. The choice for the financial ratios used in this study is based on a study by Altman (Altman et al, 1968)(Altman et al, 1977)(Altman et al, 1981) who has selected five financial ratios out of a list of 22 such ratios as providing the best classification ‘power’ for the bankruptcy prediction problem.

Bankruptcy prediction is an important method and serious topic for business. Thus, an effective prediction in time is valued priceless for business in order to evaluate risks or prevent bankruptcy. A fair amount of research has therefore focused on bankruptcy prediction. Generally speaking, there are two main groups of techniques for handling this topic: The first group is statistical techniques, such as regression analysis, correlation analysis, discriminate analysis, logit model, and probity model. The second group belongs to computational intelligence techniques such as decision trees, artificial neural networks (ANN), and support vector machines (SVM). Most researchers use one of the techniques to compare the prediction performance with other techniques for a specific data set (Odom et al, 1990)(Altman et al, 1994)(Jo et al, 1997)(Koh et al, 1999)(Shin et al, 2005)(Lensberg et al, 2006).

However, there is no single conclusion that one technique is consistently better than another for general bankruptcy prediction. During recent years researchers have paid special attention to non-parametric al and hybrid approaches. Non-parametric al techniques are suitable for
bankruptcy prediction tasks due to the specific features of financial information (i.e. non-normality and heteroskedasticity). Hybrid approaches combine several classification methods and achieve greater accuracy than individual models. The corporate bankruptcy prediction is a classical problem in the financial literature. Furthermore, bankruptcy is one of the four generic terms that are generally found in literature for corporate distress and could be defined as the condition in which a business cannot meet its debt obligations and petitions a court for either reorganization of its debts or liquidation of its assets.

The causes of business failure and bankruptcy can be pointed into: economic, financial, neglect, fraud, disaster and others. Economic factors include industry weakness and poor location, while financial factors include excessive debt and insufficient capital. For example, investors want to minimize credit risk and prevent non-profitable investments. Therefore, several authors have researched this subject in the past. Bankruptcy predictions’ impact is high for financial markets. Beaver introduced the Naïve Bayes approach in 1966 using a single variable and Altman in 1968 (Altman et al, 1968) (Altman et al, 1977) (Altman et al, 1981) proposed the use of Linear Discriminant Analysis (LDA). Since then several contributions have been made to improve the Altman’s results, using different parametric, semi parametric and non-parametric models. The use of data mining techniques such as Artificial Neural Networks (ANN), decision trees, and Support Vector Machine (SVM) for bankruptcy prediction started in the late 1980s (Kotsiantis et al, 2005) (Pompe et al, 1997) (Zhang et al, 1999) (McKee et al, 2000). Aziz and Dar (Azizi et al, 2002) carried a work of critical analysis of methodologies of corporate bankruptcies prediction models concluding that almost all models are capable of doing well their job, but the advantage of developing models based in Data Mining techniques is the future integration in Intelligent Systems. One paper (Mahmoud et al, 2012) presents a framework that evaluates a total of 9 distinct models, by comparing different algorithms (e.g. neural networks, logistic regression, svm, Bayesnet and decision trees) for year’s t, t-1, and t-2. The aim is using data mining approach for classifying the non-bankrupt and bankrupt firms.

This paper is organized as follows: The second section describes the bankruptcy prediction problem. Section three describes Genetic Algorithm. In Section four we describe our approach to the discovering rules problem the detect corporate bankruptcy. The fifth section reports on computational results evaluating the performance of the proposed system and concludes the paper.

II. Bankruptcy Prediction Problem

Bankruptcy prediction seems to be the most popular topic of the application of data mining techniques on financial data. Corporate bankruptcy causes economic damages for management, investors, creditors and employees together along with social cost. For these reasons bankruptcy prediction is an important issue in finance. Bankruptcy prediction by using financial statements data attracts its origin from the work of Altman in 1968. Altman argues that corporate failure is a long-period process and that the financial statement data should include warning signals for the imminent bankruptcy. By applying Multiple Discriminate Analysis techniques Altman developed a model for bankruptcy prediction. Since the work of Altman, many researchers developed alternative models by using statistical techniques (Mahmoud et al, 2012).
In the last years research effort has been made to build models which use data mining techniques. Lin and McClean (2001) tried to predict corporate failure by using four different methods. Two of the methods are statistical (Discriminant Analysis and Logistic Regression), whereas the remaining two methods are Machine Learning techniques (Decision Trees – C5.0 and Neural Networks). Additionally they propose a hybrid algorithm. Their sample included data about 1133 UK companies; 690 non failed companies and 106 failed companies were used as training set, whereas 289 non failed companies and 48 failed companies were used as testing set. While, no attempt was made to match failed and not failed companies. 37 financial ratios originated from balance sheet and income statements were selected as input variables. Two feature selection methods have been employed reducing the input variables to 4 by using human judgment and to 15 by using ANOVA. The authors report better results for the NNs and decision trees models for both the human judgment based and the ANOVA feature selection. Finally, the authors propose a hybrid algorithm employing weighted voting of different classifiers. Marginally better performance is reported for the hybrid model (Lin et al, 2001).

Tung et al (2004) employed a hybrid model integrating NNs and fuzzy systems. The model called “Generic Self-organizing Fuzzy Neural Network” was a rule-base consisting of IF–THEN fuzzy rules that can self-adjust the parameters of the fuzzy rules using learning algorithms derived from the NN paradigm. The main advantage of the fuzzy NN was mentioned to be its ability to model a problem by using easily comprehensible high level linguistic model instead of complex mathematical expressions. The model was applied to predict bank failures. Input variables are 9 financial variables, which have been found to be significant in previous studies. The sample contained data about 2555 non failed and 548 failed banks. 20% of the data were used as training set and 80% as testing set. To reduce Type 1 error the sample was balanced to include the equal number of failed and not failed banks. Authors report a performance of 93% when using data from the last available financial statement, 85% when using statements obtained one year prior to the last record and 75% for statements two years prior to the last available record. The model produces a set of around 50 IF-THEN fuzzy rules, which describe the interactions between the nine selected input variables and their impact on the financial health of the observed banks (Tung et al, 2004).

Shin and Lee (2002) proposed a model based on GAs. The authors stressed the fact that in contrast to the NNs, GAs can produce comprehensible rules. GAs was applied to find thresholds for one or more variables above or below which a company is considered dangerous. The model used a rule structure that contains 5 conditions each of which referred to a variable out of 9 financial ratios. The conditions were combined with the logical AND operator. The data set contained 264 failed and 264 non failed firms, whereas 9 financial ratios have been selected as input variables. 90% of the sample was used for training and 10% for validation. The general reported performance was about 80% (Shin et al, 2002).

Kim and Han (2003) built a qualitative model based on expert’s problem solving knowledge. Experts work with their subjective knowledge evaluating qualitative and quantitative facts. The model used a GA method to extract decision rules from expert’s qualitative bankruptcy predictions. The model followed the method of the experts of a Korean commercial bank. In order to predict bankruptcy the experts evaluated 6 major risk factors. In the model, a chromosome contained six segments representing a categorization of a firm according to the six risk factors. A 7th segment in the chromosome classified the firm as bankrupt or non bankrupt. The
data sample contained 772 companies, half of which were bankrupt. The experts evaluated the six risk factors for these companies. The genetic evolution process extracted 11 bankruptcy rules. Additionally, rules have been extracted by using a back propagation NN and Inductive Learning. Rules extracted with GA are reported to have better been predicting accuracy than NN and inductive learning (Dimitras et al, 1998). Dimitras et al (1998) applied RST for the aim of bankruptcy prediction. The training set contained data for 40 failed and 40 matched non failed Greek firms covering a period of five years. The testing set contained 19 failed and 19 non failed firms. A credit manager of a Greek bank selected 12 financial ratios to enter the information table and discretized the continuous values. The rough set analysis produced 54 reduces, each containing 5-7 attributes, the bank manager selected the one reduce and thus the remaining attributes were eliminated. Finally, the decision rules were derived. The results of the method have been compared with the results of discriminate analysis and logit analysis have been found to prevail (Dimitras et al, 1998).

McKee (2003) compared results obtained by using RST with actual auditors’ opinions for the purpose of bankruptcy prediction. The data sample included 146 bankrupt and 145 matched non bankrupt US companies. 11 predictive factors were chosen, 10 of which were financial ratios and 1 was a prior audit opinion. The rough set produces 87 reduces, each employing 4-6 variables and 2 reduces are selected. Two models of decision rules were developed. The results of the models were compared with actual auditors’ signaling rates and have been found almost equal. The author concludes that the models developed in this research offered no significant comparative predicting advantage over auditors’ current methodologies (McKee et al, 2003).

Beynon and Peel (2001) employed a development of RST: the Variable Precision RST. VPRST incorporated probabilistic Decision Rules and allowed partial classification by introducing a degree of confidence in classification. In contrast, previous researches efforts where the discretization of the values had been made by humans, the author mployed the FUSINTER method for the discretization purpose. The data sample contained 45 failed and 45 non failed UK industrial firms. 30 failed and 30 non failed firms form the training sample, whereas the remaining formed the holdout sample; 12 variables, 8 financial and 4 qualitative variables have been selected for the rule generation. After the re-ducts production and the selection of one of them, a set of 12 rules have been obtained. The results of VPRST were compared with results of Multiple Discriminant Analysis, Logit Analysis, Recursive Partitioning Algorithms Decision Trees and Elysee ordinal discriminant method. In the training and holdout sample VPRST outperformed some of these methods (Beynon et al, 2001).

### III. Genetic Algorithm

The idea of applying the biological principle of natural evolution to artificial systems, introduced more than four decades ago, has seen impressive growth in the past few years. Usually grouped under the term **evolutionary algorithms** or **evolutionary computation**, we find the domains of genetic algorithms, evolution strategies, evolutionary programming, and genetic programming. Such algorithms are common nowadays, having been successfully applied to numerous problems from different domains, including optimization, automatic programming, machine learning, economics, medicine, ecology, population genetics, and hardware design. In this paper we consider the evolutionary methodology known as genetic algorithms.
A genetic algorithm is an iterative procedure that involves a population of individuals, each one represented by a finite string of symbols, known as the genome, encoding a possible solution in a given problem space. This space, referred to as the search space, comprises all possible solutions to the problem at hand. Genetic algorithms are usually applied to spaces which are too large to be exhaustively searched. The symbol alphabet used is often binary, though this has been extended in recent years to include character-based encodings, real-valued encodings, tree representations, and other representations.

The standard genetic algorithm proceeds as follows: an initial population of individuals is generated at random or heuristically. Every evolutionary step, known as a generation, the individuals in the current population are decoded and evaluated according to some predefined quality criterion, referred to as the fitness, or fitness function. To form a new population (the next generation), individuals are selected according to their fitness. Many selection procedures are currently in use, one of the simplest being fitness-proportionate selection, where individuals are selected with a probability proportional to their relative fitness. This ensures that the expected number of times an individual is chosen is approximately proportional to its relative performance in the population. Thus, high-fitness (‘good’) individuals stand a better chance of ‘reproducing’, while low-fitness ones are more likely to disappear.

Selection alone cannot introduce any new individuals into the population, i.e. it cannot find new points in the search space. These are generated by genetically inspired operators, of which the most well known are crossover and mutation. Crossover is performed with probability \( p_c \) (the ‘crossover probability’ or ‘crossover rate’) between two selected individuals, called parents, by exchanging parts of their genomes (i.e. encodings) to form two new individuals, called offspring. In its simplest form, substrings are exchanged after a randomly selected crossover point. This operator tends to enable the evolutionary process to move toward ‘promising’ regions of the search space. The mutation operator is introduced to prevent premature convergence to local optima by randomly sampling new points in the search space. It is carried out by flipping bits at random, with some (usually small) probability \( p_m \). Genetic algorithms are stochastic iterative processes that are not guaranteed to converge. The termination condition may be specified as some fixed, maximal number of generations or as the attainment of an acceptable fitness level. Fig. 1 presents the standard genetic algorithm in pseudo-code format.

IV. Proposed Approach

Outline of the proposed algorithm is as follows:

Step1: Preprocessing
\[ \bar{Y} \] Normalization
\[ \bar{Y} \] Fuzzification

Step2: Generate an initial set of fuzzy if–then rules. (Initialization)

Step3: Evaluate cost of current rule-base using evaluation function.

Step4: Modify current rule-base using modify one of rules randomly and generate new rule-base.

Step5: Evaluate cost of new rule-base using evaluation function.

Step6: Admission of new rule-base with specific probability and improvement of evaluation function,

Then replace a current rule-base with new rule-base, and save best evaluation function and rule-base.
Step7: For specific iteration repeat Step4, Step5, Step6
Step8: Return best rule-base

As noted earlier data mining process is iterative and interactive process that will result in useful and valuable knowledge that is used in different places. Top Eight steps as a result of a knowledge base that includes a database and a rule base are then derived in the Inference system to use it. How to create databases and rule base in the Figure 2 is shown.

To predict and diagnose common stock by experts based on their knowledge to make decisions. But this task can be performed automatically by the computer. It is usually a system called inference system can be seen in Figure 3. The main modules in an evolving system is knowledge base, as it is more precise and more accurate decision-making process that is more acceptable which in this paper uses a highly accurate algorithm and intelligent this knowledge base is extracted. After creating a knowledge base can be used for very accurate knowledge to detect bankrupt companies from non-bankrupt company. Now we have an automated inference system based on knowledge obtained from the GA algorithm can determine with high accuracy the desired input.

```
begin GA
  g=0  \{ generation counter \}
  Initialize population P(g)
  Evaluate population P(g) \{ i.e., compute fitness values \}
  while not done do
    g=g+1
    Select \( P(g) \) from \( P(g-1) \)
    Crossover \( P(g) \)
    Mutate \( P(g) \)
    Evaluate \( P(g) \)
  end while
end GA
```

**Figure.1.** Pseudo-code of the standard genetic algorithm.

**Figure.2.** The process of creating the knowledge base (Data base & Rule base) by GA algorithm
V. Evaluation of the proposed method

The proposed algorithm assessment on data sets of stock data repository UCI (university of California at Irvin) which is used as a reference for machine learning problems. Data sets includes returns of Istanbul Stock Exchange with seven other international index; SP, DAX, FTSE, NIKKEI, BOVESPA, MSCE_EU, MSCI_EM from Jun 5, 2009 to Feb 22, 2011. In this data set repository, University of California - Irvine (UCI) is available to enable researchers to use a more accurate method for the diagnosis and prediction of bankruptcy of companies may provide and using this data set to compare the proposed method with other methods. Characteristics of these data sets can be seen in Table I.

<table>
<thead>
<tr>
<th>Data Set Characteristics:</th>
<th>Multivariate, Univariate, Time-Series</th>
<th>Number of Instances:</th>
<th>Area:</th>
<th>Business</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attribute Characteristics:</td>
<td>Real</td>
<td>Number of Attributes:</td>
<td>8</td>
<td>Date Donated</td>
</tr>
<tr>
<td>Associated Tasks:</td>
<td>Classification, Regression</td>
<td>Missing Values?</td>
<td>N/A</td>
<td>Number of Web Hits:</td>
</tr>
</tbody>
</table>

To implement and evaluate the proposed method, the programming language C++ is used. All Issues in the assessment of six popular machine learning model is used and the results obtained are compared with the proposed model (Table II).
<table>
<thead>
<tr>
<th>Proposed(GA)</th>
<th>ACO</th>
<th>MLP</th>
<th>SVM</th>
<th>Naïve Bayes</th>
<th>5-NN</th>
<th>C4.5</th>
<th>Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>92.55</td>
<td>91.88</td>
<td>85.28</td>
<td>86.78</td>
<td>87.00</td>
<td>87.22</td>
<td>84.76</td>
<td>Accuracy of Diagnosis(%)</td>
</tr>
<tr>
<td>1.45</td>
<td>1.33</td>
<td>1.6</td>
<td>2</td>
<td>1.75</td>
<td>2.35</td>
<td>2.5</td>
<td>Rules Length</td>
</tr>
</tbody>
</table>

ACKNOWLEDGMENT
The Islamic Azad University, Poldokhtar Branch have financed budget of this research plan.

References


ix. Lensberg, T, Eilifsen, A, and McKee, T, E, 2006. Bankruptcy theory development and


