Knowledge-Based Systems 24 (2011) 1348–1359

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

A novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor method

Hui-Ling Chen, Bo Yang, Gang Wang, Jie Liu, Xin Xu, Su-Jing Wang, Da-You Liu *

College of Computer Science and Technology, Jilin University, Changchun 130012, China
Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China

ARTICLE INFO

Article history:
Received 11 October 2010
Received in revised form 3 June 2011
Accepted 9 June 2011
Available online 25 June 2011

Keywords:
Fuzzy k-nearest neighbor
Parallel computing
Particle swarm optimization
Feature selection
Bankruptcy prediction

ABSTRACT

Bankruptcy prediction is one of the most important issues in financial decision-making. Constructing effective corporate bankruptcy prediction models in time is essential to make companies or banks prevent bankruptcy. This study proposes a novel bankruptcy prediction model based on an adaptive fuzzy k-nearest neighbor (FKNN) method, where the neighborhood size k and the fuzzy strength parameter m are adaptively specified by the continuous particle swarm optimization (PSO) approach. In addition to performing the parameter optimization for FKNN, PSO is also utilized to choose the most discriminative subset of features for prediction. Adaptive control parameters including time-varying acceleration coefficients (TVAC) and time-varying inertia weight (TVIW) are employed to efficiently control the local and global search ability of PSO algorithm. Moreover, both the continuous and binary PSO are implemented in parallel on a multi-core platform. The proposed bankruptcy prediction model, named PTVPSO-FKNN, is compared with five other state-of-the-art classifiers on two real-life cases. The obtained results clearly confirm the superiority of the proposed model in terms of classification accuracy, Type I error, Type II error and area under the receiver operating characteristic curve (AUC) criterion. The proposed model also demonstrates its ability to identify the most discriminative financial ratios. Additionally, the proposed model has reduced a large amount of computational time owing to its parallel implementation. Promisingly, PTVPSO-FKNN might serve as a new candidate of powerful early warning systems for bankruptcy prediction with excellent performance.

1. Introduction

Accurately identifying the potentially financial failure of companies remains a goal of many stakeholders involved. Because there is no underlying economic theory of bankruptcy, searching for more accurate bankruptcy prediction models remains the goal in the field of the bankruptcy prediction. As a matter of fact, bankruptcy prediction can be formulated as the problem of solving classification task. A fair amount of classification models has been developed for bankruptcy prediction. These models have progressed from statistical methods to the artificial intelligence (AI) approaches. A number of statistical methods such as the simple univariate analysis [1], multivariate discriminant analysis technique [2], logistic regression approach [3] and factor analysis technique [4] have been typically used for financial applications including bankruptcy prediction. Recent studies in the AI approach, such as artificial neural networks (ANN) [5–11], rough set theory [12–14], support vector machines (SVM) [15–17], k-nearest neighbor method (KNN) [18–20], Bayesian network models [21,22] and other different methods such as hybrid methods and ensemble methods [23–26] have also been successfully applied to bankruptcy prediction (see [25,26] for detail). Among these techniques, ANN has become one of the most popular techniques for the prediction of corporate bankruptcy due to its high prediction accuracy. However, a major disadvantage of ANN lies in their knowledge representation. The black box nature of ANN makes it difficult for humans to understand how the networks predict the bankruptcy.

Compared to ANN, KNN is simple, easily interpretable and can achieve acceptable accuracy rate. Albeit these advantages, the standard KNN methods place equal weights on all the selected neighbors regardless of their distances from the query point. An improvement over the standard KNN classifier is the fuzzy k-nearest neighbor classifier (FKNN) [27], which uses concepts from fuzzy logic to assign degree of membership to different classes while considering the distance of its k-nearest neighbors. It means that all the instances are assigned a membership value in each class rather than binary decision of 'bankruptcy' or 'non-bankruptcy'.

* Corresponding author at: Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012, China.
E-mail addresses: liudy@jlu.edu.cn, liudayou19420601@gmail.com (D.-Y. Liu).

0950-7051/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.knosys.2011.06.008

Points closer to the query point contributes larger value to be assigned to the membership function of their corresponding class in comparison to far away neighbors. The class with the highest membership function value is taken as the winner. The FKNN method has been frequently used for the classification of biological data [28–30], image data [31,32] and so on. Nevertheless, only few works have paid attention to using FKNN to deal with the financial problems. Bian and Mazlack [33] used FKNN as a reference classifier in their experiments in order to show the superiority of the proposed fuzzy-rough KNN method, which incorporated the rough set theory into FKNN to further improve the accuracy of bankruptcy prediction. However, they did not comprehensively investigate the neighborhood size k and the fuzzy strength parameter m, which play a significant role in improving the prediction result. This work will explore the full potential of FKNN by automatically determining k and m to exploit the maximum classification accuracy for bankruptcy prediction.

Besides choosing a good learning algorithm, feature selection is also an important issue in building the bankruptcy prediction models [25,34–37], which refers to choosing subset of attributes from the set of original attributes. The purpose of the feature selection is to identify the significant features and build a good learning model. The benefits of feature selection are threefold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data [38]. In bankruptcy prediction, genetic algorithms (GA) are usually used to select a subset of input features [39–41], to find appropriate hyper-parameter values of a predictor (for example, the kernel width and the regularization constant in the case of SVM) [35,42,43], or to determine predictor parameters (for example, multilayer perceptron weights) [44,45]. Compared with GA, PSO algorithm [46] has no crossover and mutation operators, it is simple and computationally inexpensive both in memory and runtime. Additionally, every particle adjusts their velocity and position according to the local best and global best. So that all the particles have a powerful search capability, which can help the swarm find the optimal solution. As for GA, after finding a locally optimum, it is difficult for it to find out a much better one even with a random search strategy in terms of mutation operator especially within a reasonable searching time. In this work, we will focus on exploring the PSO-based parameter optimization and feature selection approach. The continuous PSO algorithm will be employed to evolve an adaptive FKNN, where the neighborhood size k and the fuzzy strength parameter m are adaptively specified. On the other hand, the binary PSO will be used as a feature selection vehicle to identify the most informative features as well.

When dealing with the practical problems, the evolutionary-based methods such as the PSO and GA will cost a lot of computational time. There is an urgent need to improve the performance using high-performance computing techniques. For this reason, it is one of the major purposes of this paper to use a parallel environment to speed up the search and optimization process. Both the continuous and binary PSO are implemented on a multi-core platform using OpenMP (Open Multi-Processing) which is a portable, scalable model that gives programmers a simple and flexible interface for developing parallel applications for platforms [47]. The efficiency and effectiveness of the proposed bankruptcy prediction model is validated by comparing with other five state-of-the-art classification methods on two real-life cases. The experimental results demonstrate that the proposed model can not only obtain the most appropriate parameters but also show high discriminating power as a feature selection tool. Further comparison is also made between the parallel model and serial one. Based on the experiments conducted, it is inferred that the parallel model PTVPSO-FKNN can significantly reduce the computational time.

The rest of the paper is organized as follows. In Section 2, we give a brief description of the FKNN method and PSO algorithm. Section 3 proposes our model, the simultaneous optimization of relevant parameters and feature subset by the PSO approach in a parallel environment. In the next section, the detailed experimental design is presented, and Section 5 describes all the empirical results and discussion. Finally, conclusions and future work are summarized in Section 6.

2. Background materials

2.1. Fuzzy k-nearest neighbor algorithm

The k-nearest neighbor algorithm (KNN) is one of the oldest and simplest non-parametric pattern classification methods. In the KNN algorithm a class is assigned according to the most common class amongst its k-nearest neighbors. In 1985, Keller proposed a fuzzy version of KNN by incorporating the fuzzy set theory into the KNN algorithm, and named it as "fuzzy KNN classifier algorithm" (FKNN) [27]. According to his approach, rather than individual classes as in KNN, the fuzzy memberships of samples are assigned to different categories according to the following formulation:

$$u_i(x) = \frac{\sum_{j=1}^{k} u_{ij}\left(1/\|x - x_j\|^{2/(m-1)}\right)}{\sum_{j=1}^{k}\left(1/\|x - x_j\|^{2/(m-1)}\right)} \tag{1}$$

where $i = 1, 2, \ldots C$, and $j = 1, 2, \ldots, k$, with C number of classes and k number of nearest neighbors. The fuzzy strength parameter m is used to determine how heavily the distance is weighted when calculating each neighbor's contribution to the membership value, and its value is usually chosen as $m \in (1, \infty)$. $\|x - x_j\|$ is the distance between x and its jth nearest neighbor x_j. Various metrics can be chosen for $\|x - x_j\|$, such as Euclidean distance, Hamming distance, and Mahalanobis distance, among other distances. In this study, the Euclidean metric is used. u_{ij} is the membership degree of the pattern x_j from the training set to the class i, among the k nearest neighbors of x. There are two ways [27] to define u_{ij}, one way is the crisp membership, i.e., each training pattern has complete membership in their known class and non-memberships in all other classes. The other way is the constrained fuzzy membership, i.e., the k nearest neighbors of each training pattern (say x_k) are found, and the membership of x_k in each class is assigned as:

$$u_{ij}(x_k) = \begin{cases} 0.51 + (n_j/K)^{*}0.49, & \text{if } j = i \\ (n_j/K)^{*}0.49, & \text{if } j \neq i \end{cases} \tag{2}$$

The value n_j is the number of neighbors found which belong to the jth class. Note that, the memberships calculated by Eq. (2) should satisfy the following equations:

$$\sum_{l=1}^{C} \mu_{ij} = 1, \quad j = 1, 2, \ldots, n \tag{3}$$

$$0 < \sum_{j=1}^{n} u_{ij} < n \tag{4}$$

$$u_{ij} \in [0, 1] \tag{5}$$

In our experiments, we have found that the second way leads to better classification accuracy. After calculating all the memberships for a query sample, it is assigned to the class with which it has the highest membership value, i.e.,

$$C(x) = \arg\max_{i=1}^{C}(u_i(x)) \tag{6}$$

The pseudo-code of the FKNN algorithm is given below: