Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 36 (2009) 2625–2632
Expert Systems with Applications
www.elsevier.com/locate/eswa

Credit scoring algorithm based on link analysis ranking with support vector machine

Xiujuan Xu *, Chunguang Zhou, Zhe Wang

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

Abstract

Credit scoring is very important in business, especially in banks. We want to describe a person who is a good credit or a bad one by evaluating his/her credit. We systematically proposed three link analysis algorithms based on the preprocess of support vector machine, to estimate an applicant's credit so as to decide whether a bank should provide a loan to the applicant. The proposed algorithms have two major phases which are called input weighted adjustor and class by support vector machine-based models. In the first phase, we consider the link relation by link analysis and integrate the relation of applicants through their information into input vector of next phase. In the other phase, an algorithm is proposed based on general support vector machine model. A real world credit dataset is used to evaluate the performance of the proposed algorithms by 10-fold cross-validation method. It is shown that the genetic link analysis ranking methods have higher performance in terms of classification accuracy.

Keywords: Credit scoring; Link analysis ranking algorithm; Support vector machine

1. Introduction

Recently, data mining has developed rapidly over last several years, which has expanded to business and finance (Ester, Ge, Jin, & Hu, 2004; Kleinberg, Papadimitriou, & Raghavan, 1998), especially credit industry (Thomas & Edelman, 2002). For banking institutions, loans are often the primary source of credit risk. To evaluate the risk of these loans, banks use credit scoring models and credit rating to estimate default risk on a single loaner basis (Chen & Huang, 2003). However, with the intense competition of credit card issues and banks, more and more people can have credit cards and get loans from banks which have not check the applicants' credit status thoroughly.

Furthermore, many different algorithms have been proposed in pervious literature of credit scoring. The credit scoring models are developed to categorize applicants as either accepted or rejected with considering the applicants' characteristics such as age, income, and marital condition. Credit scoring is a basic binary classification task in finance. Many studies have contributed to increasing the accuracy of the classification model with various kinds of statistical tools. With the rapid growth in the credit, credit scoring models with low discriminatory power can lead to underpricing of bad and overpricing of good loans (Blochlinger & Leippold, 2006). So credit scoring need high accuracy to avoid bad debts.

Our contributions are as follows. The paper introduces a link analysis-based support vector machine (SVM) method to classify the applicants who apply for new credit cards or loans to banks. The novel approach has two phases for credit scoring. In the first phase, we present a new applicant's matrix to find the representative applicants and then find important information from the matrix by link analysis method. We compute the scoring for every applicant to distinguish good applicants and bad applicants. We call the new matrix "co-information matrix" (CIM), so that we can find the elements which are important for influence.

* Corresponding author.
E-mail addresses: xuxiujuan666@yahoo.com.cn (X. Xu), cgzhou@jlu.edu.cn (C. Zhou), wz2000@jlu.edu.cn (Z. Wang).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.01.024

We proposed the interpretation of the new model. In the second phase, we use general support vector machines (SVM) model with new input feather space.

The proposed credit scoring model is a hybrid approach using link analysis ranking techniques to preprocess samples into weighted information, and SVM techniques to build classifiers. The ranking process depends on information's values. In the SVM stage, SVM creates models to classify applicants.

The paper is organized as follows: the next section reviews some relate work about link analysis ranking and SVM in credit scoring. Sections 3 and 4 deal with the main contribution of this study. Section 3 gives a link analysis ranking model for credit data. Section 4 gives a main frame for SVM. Section 5 demonstrates the empirical evaluation including the experimental setup and the results. Conclusions are drawn in Section 6.

2. Relate works

First we describe some of the previous link analysis ranking literature (Getoor & Diehl, 2005), which has been improved and extended many aspects over several years. With the development of the Internet, information retrieve technique of search engine has developed rapidly. Link analysis technique is grown based on topology structure of the web. Link analysis ranking algorithm has exploited from web page ranking to the other fields.

Then we review some related literature in the area of support vector machine, upon which this work builds. Support vector machines are a popular data mining technique which have obtained high performance in many applications, such as credit scoring, financial time series prediction and spam categorization and so on (Martens, Baesens, Van Gestel, & Vanthienen, 2007).

Hsieh (2005) proposed a hybrid mining approach for credit scoring model. In his experiments, he specially lists the distribution of the relative importance for each input variable when using neural network.

2.1. Link analysis relate works

In the case of Web search there are two most influential hyperlink search algorithms, that is, PageRank (Brin & Page, 1998) and HITS (Kleinberg, 1998), which are related to social network. They exploit the hyperlinks of Web to rank pages according to their levels of authority. Ito Takahiko, Shimbo Masashi, Kudo Taku, and Matsumoto Yuji (2004, 2005) explored the application of kernels methods to link analysis. Borodin, Rosenthal, Roberts, and Tsaparas (2005) introduced a theoretical framework for the study of link analysis ranking algorithms. They worked with the hubs and authorities framework defined by Kleinberg (1998).

With the development of link analysis ranking, it has led to a surge of research activity in the area of information. There are a lot of workshops in many mainstream conferences, such as LinkKDD (2004, 2005, 2006), SDM (2004), SIAM (2005) and so on.

Apart from search ranking, hyperlinks are also useful for finding Web communities (Flake, Lawrence, & Giles, 2000). Beyond explicit hyperlinks on the Web, links in other contexts are useful too, for example, for ranking order individuals in a given social network in terms of a measure of their importance. Social network has become an important part in link analysis, which is the study of social entities and their interactions and relationships (Thelwall, 2006). The interactions and relationships can be described as a network or graph. From the network, we can study the properties of its structure, and the role, position and prestige of each social actor. We can also find various kinds of sub-graphs, e.g., communities formed by groups of actors. And we are more interested in social communities so as to find creditable communities from applicants, and banks can gain profit from such communities.

The technique used in link rank analysis can be extent with some latent topic models beside web graph. We would like to mine the relationship based on attributes and link structure. We are trying to construct credit communities' graphs and find important information which affects the results.

There is a fundamental assumption of link analysis that an edge between two nodes in citation graphs signifies that two nodes are in some sense related (Ito Takahiko et al., 2004). It is famous that birds of a feather flock together and things of one kind come together. Therefore, we will divide all applicants into two basis class, that is, good or bad, by their characters. Then we give another fundamental assumption in social network that two persons have the similar information of application such as income, marital condition, and then they may be possible to have similar credit. In another word, more similar information two persons have, more similar they have credit on their behavior.

2.2. SVM relate works with credit scoring

Support vector machines (Hsu, Chang, & Lin, 2003) are state-of-art data mining techniques which are popular to data classification. Recently Gold and Sollich (2005), Gold, Holub, and Sollich (2005) proposed a Bayesian method for tuning the hyperparameters of a support vector machine (SVM) classifier. They used the Nystrom approximation to the SVM kernel and their method significantly reduces the dimensionality of the space to be simulated in the hybrid Monte Carlo simulation. And then least squares support vector machine (LS-SVM) classifiers (Van Gestel et al., 2006) were applied within the Bayesian evidence framework in order to automatically infer and analyze the creditworthiness of potential corporate clients.

In credit industry, SVM has been claimed to be effective and accurate tool for credit analysis (Huang, Chen, & Wang, 2007). And various methods have been extensively employed in the credit scoring business. Piramuthu (2006)