Available online at www.sciencedirect.com

ScienceDirect

Advanced Engineering Informatics 21 (2007) 281–291

ADVANCED ENGINEERING INFORMATICS

www.elsevier.com/locate/aei

Building contextual classifiers by integrating fuzzy rule based classification technique and *k*-nn method for credit scoring

Arijit Laha

Institute for Development and Research in Banking Technology, Castle Hills, Hyderabad 500 057, India

Received 30 October 2006; accepted 22 December 2006

Abstract

Credit-risk evaluation is a very challenging and important problem in the domain of financial analysis. Many classification methods have been proposed in the literature to tackle this problem. Statistical and neural network based approaches are among the most popular paradigms. However, most of these methods produce so-called ''hard'' classifiers, those generate decisions without any accompanying confidence measure. In contrast, ''soft'' classifiers, such as those designed using fuzzy set theoretic approach; produce a measure of support for the decision (and also alternative decisions) that provides the analyst with greater insight. In this paper, we propose a method of building credit-scoring models using fuzzy rule based classifiers. First, the rule base is learned from the training data using a SOM based method. Then the fuzzy *k*-nn rule is incorporated with it to design a contextual classifier that integrates the context information from the training set for more robust and qualitatively better classification. Further, a method of seamlessly integrating business constraints into the model is also demonstrated.

Keywords: Contextual classifier; SOM; Fuzzy rule base; Fuzzy *k*-nn; Qualitative improvement; Credit scoring; Business constraints

1. Introduction

Credit scoring is a method of predicting potential risk corresponding to a credit portfolio. These models can be used by financial institutions to evaluate portfolios in terms of risk. Credit scoring tasks can be divided into two distinct types. The first type is *application scoring*, where the task is to classify credit applicants into ''good'' and ''bad'' risk groups. The data used for modelling is generally consisted of financial information and demographic information about the loan applicant. In contrast, the second type of tasks deal with existing customers and along with other information, payment history information is also used here. This is distinguished from the first type because this takes into account the customer's payment pattern on the loan and the task is called *behavioral scoring*. Recently, under BASEL II committee recommendations [16], it is increasingly becoming almost a regulatory requirement for the banks to use sophisticated credit scoring models for enhancing the efficiency of capital allocation. Data mining methods, especially pattern classification [6], using historical data, is of paramount importance in building such predictive models. In this paper, we shall focus on application scoring. However, the techniques developed here, working with appropriate data set, can be applied for behavior scoring also.

Traditionally, statistical methods are used extensively for this purpose. A survey of statistical and operation research methods for building credit and behavioral scoring models can be found in [21]. Another computational paradigm, artificial neural network (or simply neural networks (NN)) has become very popular in recent times. In contrast to the statistical methods, in NN based techniques, one need not make assumptions regarding the distribution of the data or find it through estimation techniques [9]. The NN learns the distribution implicitly from the sample data itself. This gives one great advantage

E-mail address: alaha@idrbt.ac.in

1474-0346/$ - see front matter © 2007 Elsevier Ltd. All rights reserved.
doi:10.1016/j.aei.2006.12.004

since due to "finite sample" effect, the accuracy of the estimation techniques decreases with increased dimensions of the feature space [8]. A good account of NN methods applied to various business applications including credit scoring can be found in [20]. In recent time, hybrid methods, where NN is complemented with other techniques are also being investigated. For example in [10], Hseih proposed a method of credit scoring that uses Self-organizing Map [12], K-means clustering algorithm and other NN methods. Baesens et al. [1] used Multi-layer Perceptrons along with decision trees.

Most of the statistical or NN based techniques create hard partitions of the feature space, resulting in so-called "hard" classifiers, where the classifier produces the decisions without any indication of level of confidence behind the decision. On the other hand, there is a class of classifiers, especially those incorporating fuzzy set theoretic [22] approach, termed as "soft" classifiers. These classifiers, along with the classification decision produce a confidence measure in support of it as well as alternative decisions. In other words, they have natural ability of handling uncertainty, which makes the results provided by the model more transparent and interpretable. This is extremely helpful in real life decision-making. With the help of domain experts, one can calibrate the confidence values with real life situations and an analyst using the system can make more transparent and robust decisions. Fuzzy classifiers, especially "fuzzy rule based classifiers" have been successfully used for various problem domains. A very good overview of the design techniques and applications of fuzzy classifiers can be found in [13,4].

In real life problems, the classes usually have many overlapping regions in the feature space. Every classifier encounter difficulty in correctly classifying data points in such regions. However, it is possible to address this problem, at least partially, if additional information is used for final decision-making. This additional information can be of many forms, including those from sources different from the classifier. One type of information, the *contextual information* is readily available in the sample data set used for developing the classifier. For each point to be classified, we can examine its neighborhood in the feature space to get an idea of the local class distributions around the point and integrate the information in the decision making process. One of the easiest means of doing so is employ well-known k nearest neighbor [6] rules. In its original form, k-nn classifiers find k points nearest to the point to be classified in the feature space from the training data set, and classify the point to the class from which majority of the neighbors come. In effect, the classification is based on the local context of the data point in the feature space. There are many variants of the k-nn rule. Here we consider the fuzzy k-nn rule [11], that can be easily integrated with the framework of fuzzy rule based classifiers. We call the resulting classifier fuzzy rule based k-nn (FRKNN) classifier.

In this paper, we propose a comprehensive data-driven (i.e., using learning algorithms) scheme for developing credit scoring models. The first step in this direction is to extract a good quality *fuzzy rule base* for designing a classifier. For the purpose of distinction, in this paper, we shall call this basic, non-contextual classifier fuzzy rule based (FRB) classifier. To this end we use a self-organizing map (SOM) [12] based method for fuzzy rule extraction [14,15,18] for classifier design. The fuzzy rule base is then used to design the contextual classifiers by integrating the k-nn rule for decision-making. The classifier design scheme is depicted graphically in Fig. 1, which is detailed in the following sections. Further, with the aim of developing realistic credit scoring models, we demonstrate that in the proposed scheme, various business constraints, reflecting the risk-averseness of the organization, can be incorporated very easily at the final decision making stage.

2. Building the fuzzy rule based classifier

A fuzzy rule based classifier consists of a set of fuzzy rules of the form:

R_i: **If** x_1 **is** A_{i1} **AND** x_2 **is** A_{i2} **AND** $\cdots$ **AND** x_p **is** A_{ip} **then class is** j.

Here A_{ik} is a fuzzy set used in the i-th rule and defined on the domain of attribute x_k, i.e., on the universe of the kth feature.

When a sample data point $\mathbf{x} \in \Re^p$ is presented to the system for classification, the fuzzy rules fire to produce outputs. The magnitude of the output (also known as *firing strength*) are used for deciding the class membership of the sample data $\mathbf{x}$.

For designing a fuzzy rule based classifier there are three issues that need to be addressed:

I_1: How many rules are needed?
I_2: How to generate the rules?
I_3: How to use the rules to decide a class?

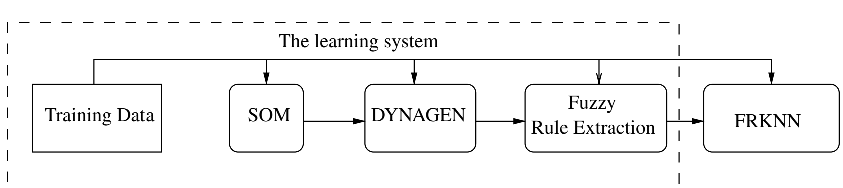

Fig. 1. Steps in developing FRKNN classifier.