International Journal of Approximate Reasoning 55 (2014) 197–210

Contents lists available at SciVerse ScienceDirect

International Journal of Approximate Reasoning

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

Incorporating logistic regression to decision-theoretic rough sets for classifications

Dun Liu [a,*], Tianrui Li [b], Decui Liang [a]

[a] School of Economics and Management, Southwest Jiaotong University, Chengdu 610031, Sichuan, PR China
[b] School of Information Science and Technology, Southwest Jiaotong University, Chengdu 610031, Sichuan, PR China

ARTICLE INFO

Article history:
Available online 16 March 2013

Keywords:
Decision-theoretic rough sets
Binary logistic analysis
Multivariate logistic regression
Decision making

ABSTRACT

Logistic regression analysis is an effective approach to the classification problem. However, it may lead to high misclassification rate in real decision procedures. Decision-Theoretic Rough Sets (DTRS) employs a three-way decision to avoid most direct misclassification. We integrate logistic regression and DTRS to provide a new classification approach. On one hand, DTRS is utilized to systematically calculate the corresponding thresholds with Bayesian decision procedure. On the other hand, logistic regression is employed to compute the conditional probability of the three-way decision. The empirical studies of corporate failure prediction and high school program choices' prediction validate the rationality and effectiveness of the proposed approach.

1. Introduction

Logistic regression analysis is a multivariate statistical method for classifications. As a representative discriminant analysis approach for classifying a set of observations into predefined classes with respect to several variables, it requires the data in an information system satisfy the conditions that the dependent variable is nonmetric and independent variables are metric [9,37,43,47]. In statistics, logistic regression is a classification method that fits data to a logistic function. It is used for predicting the outcome of a categorical criterion variable (a variable that can take on a limited number of categories) based on one or more predictor (independent or explanatory) variables. The probabilities describing the possible outcome of a single trial are modeled, as a function of explanatory variables, using a logistic function [49]. Because of no assumption regarding the distribution of the predictor variables, logistic regression is relatively robust, flexible and easily used, and it lends itself to a meaningful interpretation. In practice, logistic regression is usually used as a classifier, namely, logistic regression classifier, for probabilistic binary or multivariate classification. The criterion for selecting the category in logistic regression lies on the highest probability which is generated from a logistic function.

However, logistic regression may lead to high misclassification during the decision procedure [8]. For example, suppose an object x with three categories C_1, C_2, C_3, where $P(C_1|x) = 0.34$, $P(C_2|x) = 0.33$ and $P(C_3|x) = 0.33$. We get $x \in C_1$ according to the maximum discriminant criterion, but this criterion may lead to the misclassification for x with a high probability because of $P(C_1|x) = 0.34$ is not higher enough. For simplicity, we briefly list two drawbacks of logistic regression in real decision problems.

Problem 1: No consideration on the losses/costs of the misclassifications.

A decision is typically made under some risks and uncertainty. The misclassification may cause losses or costs. However, logistic regression does not consider losses or costs for misclassification.

* Corresponding author.
E-mail addresses: newton83@163.com (D. Liu), trli@swjtu.edu.cn (T. Li), decuiliang@126.com (D. Liang).

0888-613X/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ijar.2013.02.013

Problem 2: No consideration on the deferment decision for classification.

In logistic regression, it only considers two actions (acceptance and rejection) in a real decision problem. As well, there exist two types of misclassification scenarios in logistic regression, namely, the incorrect acceptance and incorrect rejection. This method does not consider the deferment scenario and it is actually regarded as a two-way decision [60].

To solve the two problems in logistic regression, we introduce Decision-Theoretic Rough Sets (DTRS) in our following discussions. DTRS, proposed by Yao in the early 1990s [53,54], has become a powerful decision making method and attracted more and more attention in the last 5 years. As a quantitative probabilistic extension of the qualitative classical rough set model [64], DTRS introduces Bayesian decision procedure and loss functions to rough sets. In DTRS, the pair of thresholds α and β, which are used to describe the tolerance of errors in Probabilistic Rough Sets (PRS), can be directly calculated by minimizing the decision cost with Bayesian theory. It gives a sound semantics in practical applications with minimum decision risks. Based on DTRS, Yao [60,62] further proposed the concept of a three-way decision. Obviously, the thresholds α and β can divide the universe into three pairwise disjoint regions: the positive region, boundary region and negative region. The three regions are viewed, respectively, as the regions of acceptance, rejection, and noncommitment in a ternary classification. The positive and negative regions can be used to induce rules of acceptance and rejection; whenever it is impossible to make an acceptance or a rejection decision, the boundary region can be used to induce rules of noncommitment or deferment [65]. The three-way decision is a common problem solving strategy and consistent with the idea of human decision, and it has been successfully applied to many fields, both in theories and methodologies [18,64,65], such as the extended models and their corresponding approaches on DTRS [1,10–12,14,17,21,22,25,40,41,51,61,70], attribute reduction in DTRS [13,15,30,31,58,68], applications on DTRS [2,3,6,7,19,20,24,26,28,29,33,42,46,50,52,66,67,69]. Although DTRS has achieved lots of achievements in many domains, it has met some big challenges yet. One of them is how to calculate both the conditional probability and the thresholds in DTRS.

For calculating the conditional probability, Yao and Zhou [61] proposed a naive Bayesian DTRS model. The conditional probability is estimated based on the Bayes' theorem and the naive probabilistic independence assumption. In order to compute the thresholds α and β in DTRS, Herbert and Yao [10–12] introduced a game-theoretic approach to DTRS for learning optimal parameter values. Measures of classification ability are interpreted as players in a game, each with a goal of optimizing its value by increasing or decreasing the size of the classification regions, and the optimal parameter values are generated by the Nash equilibrium in game theory. Li and Zhou [14] argued that the thresholds of probabilistic inclusion are calculated based on the minimization of risk cost under the optimistic decision, the pessimistic decision, or the equable decision. Deng and Yao [4] utilized an information-theoretic interpretation of thresholds in PRS. In this paper, we provide an alternative method for computing the conditional probability required by DTRS, which leads to an effective way to estimate the required conditional probability. We also generalize DTRS to the multiple category classification. The results enable us to apply DTRS in solving real-world classification problem.

In this paper, we propose an integrated classification approach by using of logistic regression and DTRS. On the one hand, DTRS considers the costs and the deferment decision in a real problem. In the view of semantics, DTRS is utilized to systematically calculate the corresponding thresholds by considering cost or loss. DTRS can reasonably explain the thresholds of logistic regression and supply the deferment strategy for the classification. On the other hand, as we stated, the three-way decision in DTRS depends on a pair of thresholds and conditional probability [19]. Observed by the continuous and discrete data may coexist in information systems in real applications, logistic regression provides a way to compute the conditional probability in this situation, and solves the mentioned challenge of DTRS.

The remainder of this paper is organized as follows: Section 2 provides the basic concepts of logistic regression and DTRS. In Section 3, we combine the logistic regression and DTRS, and propose two novel integrated classification models to solve the binary misclassification problem and multiple classification problem, respectively. Then, the case studies of corporate failure prediction and high school program choices' prediction are given to illustrate our approaches in Section 4. Section 5 concludes the paper and outlines the future work.

2. Preliminaries

Basic concepts, notations and results of the logistic regression analysis and DTRS are briefly reviewed in this section [5,9,16,18,20,27,34–36,39,38,44,45,48,49,60,71,72].

2.1. Logistic regression analysis

This subsection introduces the logistic regression analysis to calculate the conditional probability for objects in an information table. As we known, logistic regression can be bi- or multinomial. Binary or binomial logistic regression refers to the instance in which the observed outcome can have only two possible types (e.g., "dead" vs. "alive", "success" vs. "failure", or "yes" vs. "no"). Multivariate or multinomial logistic regression refers to cases where the outcome can have three or more possible types (e.g., "better" vs. "no change" vs. "worse") [49]. As a common logistic regression analysis approach, binary logistic regression and multinomial logistic regression methods are utilized to deal with the classification problem because they can directly compute the probability of occurrence of an event.