Expert Systems with Applications 36 (2009) 5900–5908

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Evolutionary-based feature selection approaches with new criteria for data mining: A case study of credit approval data

Chia-Ming Wang [a], Yin-Fu Huang [b,c,*]

[a] Graduate School of Engineering Science and Technology, National Yunlin University of Science and Technology, 123 University Road, Section 3, Touliou, Yunlin 640, Taiwan, ROC
[b] Graduate School of Computer Science and Information Engineering, National Yunlin University of Science and Technology, 123 University Road, Section 3, Touliou, Yunlin 640, Taiwan, ROC
[c] Department of Computer and Communication Engineering, National Yunlin University of Science and Technology, 123 University Road, Section 3, Touliou, Yunlin 640, Taiwan, ROC

ARTICLE INFO

Keywords:
Data Mining
Evolutionary algorithm
Feature selection
Multi-objective optimization

ABSTRACT

In this paper, the feature selection problem was formulated as a multi-objective optimization problem, and new criteria were proposed to fulfill the goal. Foremost, data were pre-processed with missing value replacement scheme, re-sampling procedure, data type transformation procedure, and min-max normalization procedure. After that a wide variety of classifiers and feature selection methods were conducted and evaluated. Finally, the paper presented comprehensive experiments to show the relative performance of the classification tasks. The experimental results revealed the success of proposed methods in credit approval data. In addition, the numeric results also provide guides in selection of feature selection methods and classifiers in the knowledge discovery process.

1. Introduction

Recently, data mining or knowledge discovery in databases (KDD) has emerged as a very active, evolving area in information technology. Hundreds of novel mining algorithms and new applications such as medicine, business, and engineering have been proposed in the last decade. The aim of data mining is to extract knowledge from data (i.e., to help human finding and interpreting the 'hidden information' in massive raw data). The information and knowledge mined from the large quantities must be meaningful enough to lead to some advantages, usually economic advantages (Witten & Frank, 2005).

A credit scoring technique is the set of decision models and their fundamental techniques assist lenders in the granting of consumer credit (Thomas, 2000). It has been extensively used for credit admission evaluation in recent years. The basic principle of credit scoring is based on the analysis of the past performance of consumers to predict the credit score of those who will be assessed. In fact, the essential operations and philosophy are similar to the knowledge discovery process. Researchers have developed a variety of parametric statistical models such as LDA and logistic regression models (Desai, Crook, & Overstreet, 1996) for credit scoring. Nevertheless, assumptions of the underlying probability distribution are essential part of these methods. Moreover, those methods also assume linear relationships between attributes. These restrictions or shortages decrease the predictive accuracy of the credit scoring models and prevent their success.

In this paper, we applied meta-heuristic search techniques to find approximations of Pareto optimal set for the feature selection problem. Moreover, we proposed two new objectives for this combination optimization problem. Some pre-processing steps were conducted before the knowledge discovery process.

The primary contributions of the paper are as follows:

1. Since the feature selection problem could be considered as a combination optimization problem, the paper proposed new criteria for single/multiple objective evolutionary feature selection. The paper presented comprehensive experiments to show the relative performance of the classification tasks in the knowledge discovery process.
2. The results of an empirical study presented the relative performance of five different feature selection techniques. The results show:
   (a) New criteria with evolutionary algorithm outperform other feature selection methods.
   (b) K-nearest neighbor classifier usually produces poor performance no matter what performance measure is used.

The remainder of this paper is organized as follows. Section 2 described the workflow of the knowledge discovery process. How we preprocess data instances were described precisely in the section.

* Corresponding author. Tel.: +886 5 5342601x4314; fax: +886 5 5312063.
E-mail addresses: wang.chia.ming@gmail.com (C.-M. Wang), huangyf@yuntech.edu.tw (Y.-F. Huang).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.07.026

Section 3 introduced the feature selection problem and proposed solutions. The new objectives for single/multiple objective optimization were proposed in the section. In Section 4, we presented experimental setting and results. Finally, we concluded in Section 5.

2. Learning system

2.1. Workflow

Data pre-processing is always the first step (even the most important one) in the data mining workflow. Without getting to know data carefully in advance, the classification task could be misleading. First, the whole data sets were dealt with missing value replacement scheme. Then, a re-sampling procedure, including up-sampling scheme and down-sampling scheme, was performed for tackling a data imbalance issue. Finally, nominal attributes of data instances were transformed into numeric attributes, then normalized by a min–max normalization procedure, and finally fed into a feature selection module sequentially.

After going through the pre-processing procedures and feature selection module, the whole data is randomly divided into five divisions of equal size. The class in each division is represented in nearly the same proportion as that in the whole data set. Each division is held out in turn and the remaining four-fifths are directly fed into the classifiers. Thus, classifiers are executed 5 times on different training sets. This *k*-fold cross validation procedure could minimize the impact of data dependency and prevent the over-fitting problem (Hsu, Chang, & Lin, 2003). The detail workflow is shown in Fig. 1.

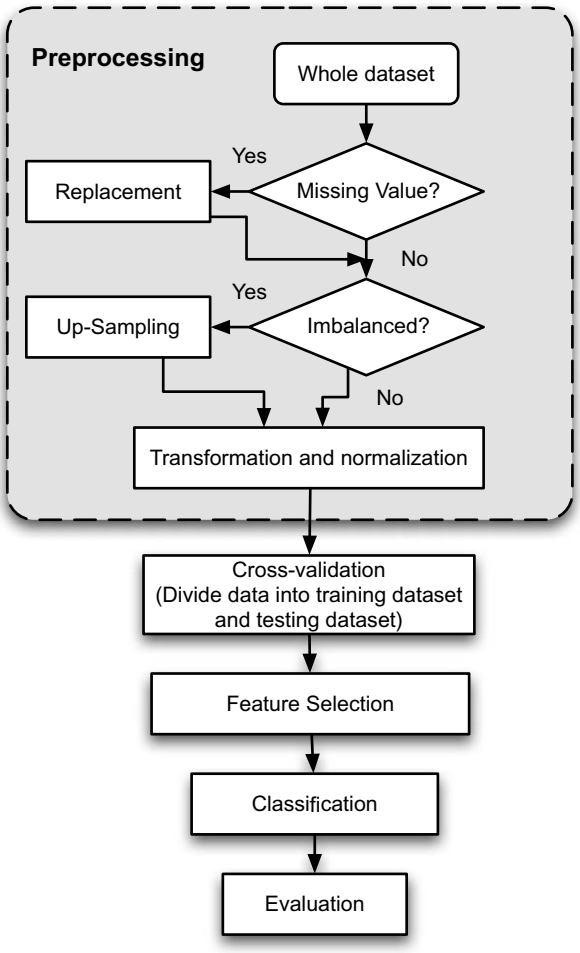

Fig. 1. Workflow of knowledge discovery.

2.2. Pre-processing

In this section, we explain how the data instances are pre-processed. The whole data sets were dealt with missing value replacement scheme, re-sampling procedure, data transformation procedure, and min-max normalization procedure sequentially.

2.2.1. Missing value replacement

Since most data sets encountered in practice contain missing values and most learning schemes lack for ability to handle these data sets, we have replaced missing values with the average or mode of attributes depending on their attribute types; i.e., numerical or categorical ones. Indeed, it seems to be convenient alternatives to remove all of these instances as long as the quantities of data are not too many.

2.2.2. Re-sampling

Recently, the class imbalance problem has been an interesting topic in machine learning and data mining community (Weiss & Provost, 2001). When classes are imbalanced, it would cause seriously negative effects on the classification performance; i.e., the overall error rate (Drummond & Holte, 2003). Most practical classifiers not designed for cost-sensitivity do much better on majority classes since they have a bias towards generality. However, in the worst case, classifiers do nothing on minority classes and predict the entire sample to majority ones. Cost-sensitive learning and re-sampling are two general methods to handle this problem, although there is no consistent winner. Since most algorithms are not cost-sensitive inherently, we adopt a re-sample approach to deal with this problem.

2.2.3. Data transformation and normalization

Some machine learning schemes such as neural network and SVM require that each data instance is represented as a vector of real numbers. Therefore, we have to convert the nominal attributes into numeric data before feeding into classifiers. Instead of using a single-number to represent a nominal attribute, we used *k* numbers to represent all *k* distinct nominal values of an attribute. That is, only one of the *k* numbers is one, and others are all zero. Apparently, this coding uses more numeric attributes to represent one nominal attribute, but it might be more stable than using a single-number if distinct values of an attribute are not too many (Hsu et al., 2003).

In order to prevent attributes with large numeric ranges dominate those with small numeric ranges, data instances are rescaled between 0 and 1 using min–max normalization procedure. The min-max normalization procedure performs a linear transformation of the original input range into a new specified range. The old minimum min_old is mapped to the new minimum min_new (i.e., 0) and max_old is mapped to max_new (i.e., 1), as shown in Eq. (1).

$$\text{New\_value} = \frac{\text{original\_value} - \text{min\_old}}{\text{max\_old} - \text{min\_old}}(\text{max\_new} - \text{min\_new}) + \text{min\_new} \quad (1)$$

3. Feature selection

The feature selection module in the knowledge discovery process aims to select the most relevant features. There are three regular approaches for feature selection – filter-based, wrapper-based and embedded-based ones (Huang, 2003). The filter approach