A new approach for rule extraction of expert system based on SVM

CrossMark

Ai Li, Guo Chen *

College of Civil Aviation, Nanjing University of Aeronautics & Astronautics, Nanjing 210016, PR China

ARTICLE INFO

ABSTRACT

Based on the SVM's excellent generalization performance, a new approach is proposed to extract knowledge rules from Support Vector Clustering (SVC). In this method, the first step is to choose the features of the sample data by using Genetic Algorithm for improving the comprehensibility of the knowledge rules. Then the SVC algorithm is adopted to obtain the Clustering Distribution Matrix of the sample data whose features have been chosen. Finally, hyper-rectangle rules are constructed using the Clustering Distribution Matrix. To make the rules more concise, and easier to explain, hyper-rectangle rules are simplified further by using rules combinations, dimension reduction and interval extension. In addition, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm is adopted to resample fault samples in order to solve the serious imbalance problem of samples. The UCI datasets are used to validate the new method proposed in this paper, the results compared with other rules extraction methods show that the new approach is more effective. The new method is used to extract knowledge rules for aero-engine oil monitoring expert system, and the results show that the new method can effectively extract knowledge rules for expert system, and break through the bottleneck in expert system knowledge dynamic acquisition.

1. Introduction

At present, knowledge acquisition through data mining [1,2] occurs mainly through machine learning or statistics. Correlation analyses [3], artificial neural networks [4], rough sets [5], and decision trees [6] are extensively employed for data mining. If data mining is applied to an expert system and if the knowledge rules are extracted automatically from real data, then the intelligence level and knowledge acquisition ability of the expert system will be greatly improved.

In recent years, the Support Vector Machine (SVM) [7] has become an emerging classification technology in data mining. The SVM can approximate any continuous bounded nonlinear function because of the perfect general-ization theory and strong nonlinear mapping ability. The SVM has several advantages over the neural network, such as better generalization ability, no local minimum problem, the ability to automatically construct the learning machine, no dimension curse, and the ability to deal with small samples. These advantages have caused data mining technology based on SVM to receive the attention of researchers worldwide. Furthermore, a number of promising SVM rule extraction algorithms published to date [8–14] are not only simple but also broadly applicable. Nunez et al. [9] introduced a rule extraction approach based on the SVM, in which K-means clustering is used to obtain clustering centers, which are then combined with support vectors (SVs) to define ellipsoid rules. Finally, the "if-then" rules can be obtained when the ellipsoid rules are mapped to the input space. However, the generated ellipsoid rules seriously overlap. In addition, the solution quality of K-means strongly depends on the initial values for the centers, and it is difficult to control the quantity and quality

* Corresponding author. Tel./fax: +86 025 84891850.
  E-mail address: cgzyx@263.net (G. Chen).

of the obtained rules. In a similar study, Zhang et al. [10] introduced the hyper-rectangle rule extraction (HRE) algorithm to extract rules from the trained SVM. The authors used the Support Vector Clustering (SVC) algorithm to find prototype vectors for each class, and then used those vectors with the SVs to generate hyper-rectangles. A nested generalized exemplar algorithm is utilized to first construct small hyper-rectangles around the prototypes, which are then grown incrementally until the stopping criteria based on a user-defined minimum confidence threshold (MCT) or minimum support threshold (MST) are met. If-then rules are then generated by projecting these hyper-rectangles onto coordinate axes. The published results for this method show that the rules provide good accuracy. However, all the features are present as antecedents of these rules. This limits their explanation capability, since no indication is given about the most important features for the classification.

Based on the aforementioned limitations, here, a new method is proposed to extract knowledge rules from SVC. The first step in this method is to choose the features of the sample dataset using a Genetic Algorithm (GA) for improving the comprehensibility of the knowledge rules. The next step is to map the chosen features of the training samples into a high-dimensional feature space to get optimal separating hyper-planes and SVs. Finally, the hyper-rectangles are constructed using the Clustering Distribution Matrix of the data obtained by the SVC, and the if-then rules are generated by projecting these hyper-rectangles onto coordinate axes. In order to make the rules more concise and easier to explain, hyper-rectangle rules are further simplified using a combination of rules, dimension reduction, and interval extension. In addition, the SMOTE (Synthetic Minority Over-sampling Technique) algorithm is adopted to resample fault samples in order to solve the serious imbalance problem of samples. Experimental results show that it is easy to control the number and the support degree of the generated rules; feature selection and simplification of rules can greatly improve their explanation capability.

Spectral oil diagnosis expert system is the advanced stage of aero-engine wear fault diagnosis. At present, some oil monitoring expert systems have been developed, such as, the advanced rapid analysis system PFALink developed by the United States Mobil oil company, lubricating oil analysis expert system Lube Analyst and Atlas developed by the United States and Canada. But these software only provides a framework and management system, and the

users need to develop its core knowledge base and provide the monitored wear element threshold value. In the intelligent diagnosis expert system, these problems, such as weak knowledge acquisition, hard knowledge updating and poor knowledge adaptability, still did not get effective to be overcome. The expert system knowledge acquisition is basically by means of the mechanical learning methods based on the experiences. The knowledge is hard to update and the rules exist serious problems such as inconsistent, redundancy, and combination explosion. Therefore, in this paper, the new method is applied to the knowledge acquisition of aero-engine spectral oil diagnosis expert system. Experimental results to real dataset show the effectiveness and the correctness of the new method.

2. Knowledge rules extracting method based on GA_SVC

The rule extraction process includes data preprocessing, SVC, hyper-rectangle rule extraction and rule simplification. The entire rule extraction procedure is shown in Fig. 1.

2.1. Data preprocessing

2.1.1. Dalancing to unbalance data

In data mining experiments, the datasets are usually assumed to balance distribution, which is the number of various types of samples is almost the same, while it is almost non-existent in the real. In many real datasets, the number of class with different label is unequal. These datasets are called unbalanced datasets. Usually, the minority class samples will be taken out as noise so that no rules about the minority class can be extracted. Therefore, in order to extract rules of various types of samples completely, and improve the recognition rate of the rules, the first step is to preprocess the unbalance data into balance data before rules extraction.

In this paper, we resample fault samples by using Synthetic Minority Over-sampling Technique (SMOTE) which is the typical sampling algorithm. SMOTE [15] algorithm is an over-sampling method put forward by Chawla. In order to make the dataset be equilibrium, the main concept of the method is to use k neighbor method and linear interpolation method to insert new samples according to certain rules between the two closer samples of minority class. In Fig. 2, a two dimensional example $\{\mathbf{X} = (x_1, x_2)\}$ is enlarged by using SMOTE over-sampling method. It can be seen from Fig. 2 that the new re-sampling samples focus
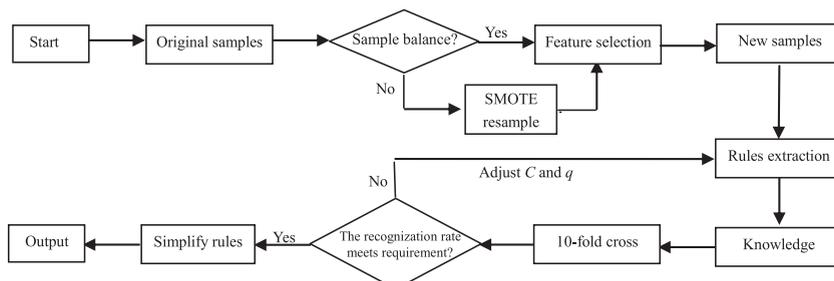

Fig. 1. Rules extraction procedure.