Expert Systems With Applications 64 (2016) 239–246

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A new support vector data description method for machinery fault diagnosis with unbalanced datasets

Lixiang Duan, Mengyun Xie, Tangbo Bai, Jinjiang Wang*

School of Mechanical and Transportation Engineering, China University of Petroleum, Beijing 102249, China

ARTICLE INFO

Article history:
Received 25 February 2016
Revised 6 June 2016
Accepted 26 July 2016
Available online 27 July 2016

Keywords:
Fault diagnosis
Unbalanced datasets
Support vector data description
Binary tree
Mahalanobis distance

ABSTRACT

In machinery fault diagnosis area, the obtained data samples under faulty conditions are usually far less than those under normal condition, resulting in unbalanced dataset issue. The commonly used machine learning techniques including Neural Network, Support Vector Machine, and Fuzzy C-Means, etc. are subject to high misclassification with unbalanced datasets. On the other hand, Support Vector Data Description is suitable for unbalanced datasets, but it is limited for only two class classification. To address the aforementioned issues, Support Vector Data Description based machine learning model is formulated with Binary Tree for multi-classification problems (e.g. multi fault classification or fault severity recognition, etc.) in machinery fault diagnosis. The binary tree structure of multiple clusters is firstly drawn based on the order of cluster-to-cluster distances calculated by Mahalanobis distance. Support Vector Data Description model is then applied to Binary Tree structure from top to bottom for classification. The parameters of Support Vector Data Description are optimized by Particle Swarm Optimization algorithm taking the recognition accuracy as objective function. The effectiveness of presented method is validated in the rotor unbalance severity classification, and the presented method yields higher classification accuracy comparing with conventional models.

1. Introduction

Machine learning techniques have been widely investigated for intelligent machinery fault diagnosis (McBain & Timusk, 2009). An improved neural network algorithm was investigated for incipient locomotive roller bearings fault diagnosis (Lei, He, Zi, & Hu, 2008). A set of individual neural networks based on structured genetic algorithm were constructed for rotating machine fault diagnosis (Chen & Chen, 2011). Support Vector Machine (SVM) combined with Ant Colony algorithm was applied for rotating machinery fault diagnosis (Zhang, Chen, Wang, & Chen, 2015). Fuzzy C-Means (FCM) was also applied for motor rotor fault diagnosis (Cao, Li, & Jiang, 2013). These commonly used machine learning techniques, such as Neural Networks, SVM, and FCM have presented great advantages of competitive classification accuracies when the dataset numbers of different classes are almost even. However, when the datasets among different classes are unbalanced, these machine learning techniques usually suffer from low classification accuracies.

In practice, machinery usually operates under normal condition. The faults seldom occur, but the fault types are various. The data samples under faulty conditions are usually expensive to obtain, and even not available (Chen, Zhang, Wang, & Li, 2015). Data scarcity under machinery faulty conditions causes the data amounts acquired in the faulty conditions are far less than those of normal condition, resulting in unbalanced dataset issue (McBain & Timusk, 2011). When the dataset is unbalanced, it is difficult to use the data information samples to construct an accurate classifier for conventional machine learning algorithms. The performance of these classifiers are unsatisfactory especially when the number of abnormal samples is far less than that of normal samples. Abnormal samples are easy to be neglected and submerged by the majority class, which may cause fault misclassification, eventually leading to equipment breakdown, disastrous accidents and economic losses.

Many research efforts have been put to address the unbalanced dataset issue from different perspectives, which can be divided into improved algorithms and data preprocessing techniques, etc. The improved algorithms, such as cost-sensitive learning, ensemble learning, probability density change, etc. are beneficial for minority class samples to increase the accuracies. For example, cost-sensitive classification considers different costs of varying fault types, using cost matrix to minimize the cost of error. Many

* Corresponding author.
E-mail addresses: duanlx@cup.edu.cn (L. Duan), xxmmyy1992@163.com (M. Xie), twzone@163.com (T. Bai), jwang@cup.edu.cn (J. Wang).

http://dx.doi.org/10.1016/j.eswa.2016.07.039
0957-4174/

studies have applied cost-sensitive learning to Naïve Bayes (Sheng, Ling, & Yang, 2005) and decision trees (Elkan, 2001), however, cost-sensitive classification faces two challenges: (1) misclassification cost could not be determined; (2) it is usually difficult to find an effective method of evaluating the performance of a cost-sensitive classifier. Some research efforts have concentrated on ensemble learning, Yan, Liu, Jin, and Hauptmann, 2003 used ensample learning to address unbalanced datasets. Liu, Wu, and Zhou, 2009 combined ensemble methods with under-sampling. However, the performance of such method relies on basic learners.

From the data preprocessing perspective, resampling techniques including over-sampling and sub-sampling have been investigated to reduce the unbalanced degree. Estabrooks, Jo, and Japkowicz, 2004 combined different expressions of resampling approach to analyze unbalanced problem. A simple over-sampling method is to copy samples of minority class to increase the sample number in minority class. However, the disadvantage of this method is that no new knowledge is increased which leads to over fitting. For the sub-sampling, it reduces the scale of majority class by removing the samples of majority class. However, it may cause significant data information lost.

On the contrary, Support Vector Data Description (SVDD) describes the data by a spherically shaped decision boundary, and has been applied for outlier or novelty detection in one-class classification (Tax & Duin, 1999; Tax, Ypma, & Duin, 1999). Recently, it arises much research interest in machinery fault diagnosis because of its advantages of flexible description and complimentary classification accuracy. SVDD is combined with SVM to solve the data description problem with negative samples (Wang, Zhao, Weng, & Zhang, 2015). The decision boundary of SVDD is set by the threshold of cluster distance to improve the overall performance of SVDD (Ramirez & Allende, 2012). The application of SVDD was also extended to multi-classification problem. An improved SVDD method (Hao, Chiang, & Lin, 2009) was proposed to find the maximal-margin of spherical structure and to construct spheres for each class at one time. Lee and Lee, 2007 proposed the classifier based on the Bayesian optimal decision theory to deal with multi-class classification. However, decision making function did not consider the volume of the hyper-sphere. When the number of the dataset is large, it may cause the problem of overlapping and some data may be categorized into two classes or even more.

To address the aforementioned issues, a new method integrating SVDD with Binary Tree based on Mahalanobis distance is proposed for multi-classification issue in machinery fault diagnosis. The binary tree structure of multiple clusters is firstly drawn based on the order of cluster-to-cluster distances calculated by Mahalanobis distance. SVDD model is then applied to the Binary Tree structure from top to bottom for classification. The parameters of SVDD are optimized by PSO algorithm taking the recognition accuracy as objective function. The effectiveness of presented method is validated in the unbalance severity classification of rotor, and the presented method yields higher classification accuracy comparing with the conventional models.

The intellectual merits of this study rest on three folds. 1) A new support vector data description method integrating with Binary Tree is presented for multi-classification in machinery fault diagnosis. 2) A separability measure based on Mahalanobis distance is proposed to construct Binary Tree. 3) The parameters of SVDD are optimized by PSO algorithm to eliminate the error caused by manually selection. The rest of this paper is organized as follows. The theoretical background of SVDD, Mahalanobis distance and PSO are introduced in Section 2. The proposed method BT-PSO-SVDD is discussed in Section 3. In Section 4, the effectiveness of presented BT-PSO-SVDD method is experimentally demonstrated in rotor unbalance severity recognition, including the comparison results with other approaches. In Section 5, the discussions are unfolded. In Section 6, the concluding remarks are drawn.

2. Theoretical background

2.1. Support vector data description

Inspired by SVM, Tax and Duin, 1999 proposed a new method-Support Vector Data Description, which can be applied to outlier or novelty detection. The one-class classier SVDD is the transformation of SVM. The main purpose of this method is to replace the hyper-plane of SVM by a hyper-sphere which can describe the data by a spherically shaped decision boundary. Once the hyper-sphere has been identified, it can be used as the classifier to predict the unknown sample and evaluate whether the unknown sample belongs to that class.

The theory of SVDD can be described as follows:

$$\min_{R,a,\xi_i}(R, a, \xi_i) = R^2 + C\sum_{i=1}\xi_i \tag{1}$$

where a is the center of the hyper-sphere and R is the radius of the hyper-sphere. C is the penalty factor which controls the spherical volume and error. Almost all objects are within the sphere:

$$s.t.\|x_i - a\|^2 \le R^2 + \xi_i, \quad \xi_i \ge 0\forall i \tag{2}$$

ξ_i is a slack variable for allowable error range. To find the value of Eq. (1) is difficult, Lagrange function is constructed, α_i and β_i are Lagrange multipliers, and the formulation can be transformed as follows:

$$L(a, R, \alpha_i, \beta_i, \xi_i) = R^2 + C\sum_i \xi_i - \sum_{i=1}^{n}\alpha_i(R^2 + \xi_i - \|x_i - a\|^2) - \sum_{i=1}^{n}\beta_i\xi_i \tag{3}$$

$$s.t.\sum_i \alpha_i = 1, 0 \le \alpha_i \le C \tag{4}$$

In Eq. (3), calculate the derivative of a, R, ξ_i in which their values are set as 0.

$$\frac{\partial L}{\partial a} = 0 \Rightarrow a = \sum_{i=1}^{n}\alpha_i x_i \tag{5}$$

$$\frac{\partial L}{\partial R} = 0 \Rightarrow \sum_{i=1}^{n}\alpha_i = 1 \tag{6}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \beta_i = C - \alpha_i \tag{7}$$

Bring Eqs. (5)–(7) into Eq. (3), and Eq. (3) can be transformed as:

$$L = \sum_{i=1}^{n}\alpha_i\left\|x_i - \sum_{j=1}^{n}\alpha_j x_j\right\|^2 \tag{8}$$

The dual form of the optimization problem can be described as:

$$\max\sum_{i=1}^{n}\alpha_i(x_i, x_i) - \sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j(x_i, x_j) \tag{9}$$

$$s.t.\sum_{i=1}^{n}\alpha_i = 1, 0 \le \alpha_i \le C, i = 1, 2, \cdots n \tag{10}$$