Journal of Process Control 64 (2018) 62–70

Contents lists available at ScienceDirect

Journal of Process Control

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

Weighted random forests for fault classification in industrial processes with hierarchical clustering model selection

Yue Liu, Zhiqiang Ge*

State Key Laboratory of Industrial Control Technology, Institute of Industrial Process Control, College of Control Science and Engineering, Zhejiang University, Hangzhou 310027, Zhejiang, PR China

ARTICLE INFO

Article history:
Received 11 August 2017
Received in revised form 9 February 2018
Accepted 10 February 2018

Keywords:
Hierarchical clustering
Weighted random forests
Ensemble learning
Model selection
Fault classification

ABSTRACT

In this paper, a hierarchical clustering selection based weighted random forests scheme is proposed for fault classification in complex industrial processes. Model diversity and the strength of each model are deemed to be two key issues for the performance of ensemble learning method. To improve the diversity between classification trees and the performance of individual classification trees in random forests, the hierarchical clustering method is applied for offline model selection in random forests, which can simultaneously reduce the online fault classification complexity. Meanwhile, the weighted voting rule is used in random forests instead of majority voting, in order to boost the good performance models and weaken the bad ones. Detailed comparative studies between proposed method and conventional methods have been carried out through the Tennessee Eastman (TE) benchmark process.

1. Introduction

Process monitoring has become increasingly important in industrial processes. As industrial systems have become more highly integrated and complex, first-principle model and knowledge-based approaches are quite difficult to be applied for process monitoring [1]. At the same time, large quantities of data have been collected with the widespread use of distributed control systems (DCS) in industrial processes. Coupled with the improvement of storage capacity and the computational speed of computer, data-driven approaches become the most popular techniques for process monitoring, which includes fault detection, fault identification, fault classification and fault diagnosis [2,3]. Over the past decades, a number of multivariable statistical approaches have been proposed for fault detection, which are not typically used for fault classification, such as Principal component analysis (PCA) [4,5], partial least square (PLS) [6,7], etc. [8–11]. However, fault classification, aiming to determine the type of the detected fault, is essential to the elimination of fault and process recovery. Therefore, fast and accurate fault classification is significant for a well-designed process monitoring system.

In recent years, numerous machine learning methods have been employed for the purpose of fault classification, which can be considered as a multi-class classification problem. For example, Chiang et al. proposed fisher discriminant analysis (FDA) for fault diagnosis [12]. Zhong et al. proposed a semi-supervised FDA model for fault classification with only a few labeled data samples [13]. Yuan and Chu proposed a multi-class support vector machines (SVM) method for fault diagnosis of turbo-pump rotor [14]. Zhang et al. proposed an artificial neural network (ANN) approach for transformer fault diagnosis [15]. For the fault classification problem in complex industrial processes, it is difficult to design a single classifier to achieve the desired performance under different circumstances. For example, SVM method is not good at dealing with multi-class problem and is sensitive to the missing data. ANN method is difficult to interpret and has many parameters need to be tuned. To solve this problem, ensemble learning methods have been proposed in several research works for fault classification [16–21], which are able to overcome the weakness of the single classifier and simultaneously improve the performance of fault classification. Among those ensemble learning methods that have been used for fault classification, random forests is a popular one, which has been used in various areas since it was proposed by Breiman [22]. Random forests is a tree-structured ensemble learning method, which consists of a large number of trees and then votes for the most popular class. Due to its two characteristics: boostrap sampling of training set and random selection of features at each node of the tree, random forests has superior performance over other ensemble machine learning methods in both classification and regression.

Traditional ensemble learning methods combine all classifiers to achieve the final result. However, some research works indicate that selective ensemble, which selects a part of the classifiers

* Corresponding author.
E-mail address: gezhiqiang@zju.edu.cn (Z. Ge).

https://doi.org/10.1016/j.jprocont.2018.02.005
0959-1524/

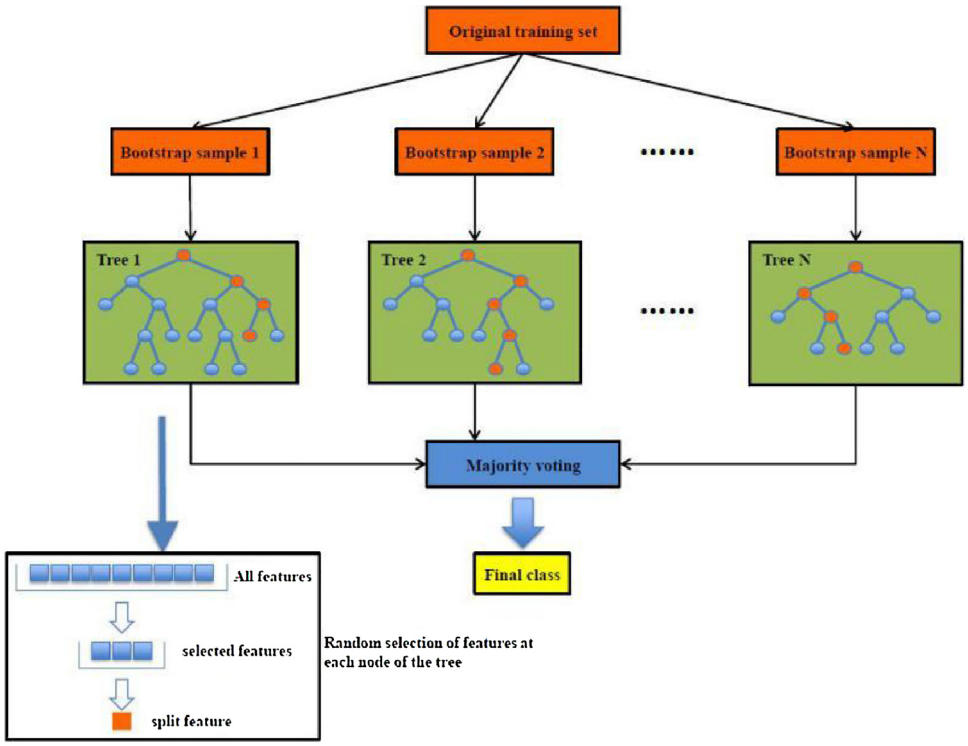

Fig. 1. Typical structure of random forests.

to make up an ensemble, is better than all ensemble. For example, Zhou et al. proposed a selective ensemble method GASEN and applied it to the neural network ensemble, which uses less neural networks but achieves stronger generalization ability [23]. Another method GASEN-b was also proposed by Zhou et al., which proves the effectiveness of selective ensemble of decision trees [24]. Therefore, to improve the online classification performance of random forests, a certain number of decision trees can be selected from the trained trees. To make an effective selection, two key issues, which are the diversity between classifiers and the strength of each classifier, should be paid attention. In this paper, hierarchical clustering method which makes the two issues mentioned above well-balanced is proposed for decision tree selection in random forests. Hierarchical clustering is a connectivity-based clustering method, in which objects in same cluster are more similar to each other than those in different clusters [25]. After hierarchical clustering, the tree with best performance in each cluster is selected. Meanwhile, weighted voting fusion strategy is used in random forests instead of majority voting strategy, which can boost the good performance of selected trees and weaken the bad ones.

The main contribution of this paper can be summarized as follows. Firstly, a multitude of decision trees are constructed to form the random forests by bootstrap sampling and random selection of features at each node of the tree. Secondly, the performance of each decision tree is evaluated by validation set and then the dissimilarity between the trees is calculated, which is used as the distance in hierarchical clustering method for decision tree selection in random forests. Thirdly, online fault classification is carried out through the new random forests model with decision trees selection and weighted voting fusion strategy.

The rest of this paper is organized as follows. Section 2 provides a review of random forests. Hierarchical clustering method is introduced in Section 3. Section 4 illustrates the framework of hierarchical clustering selection based weighted random forests method and its application for fault classification, In Section 5, the performance of the proposed method is evaluated through the Tennessee Eastman (TE) benchmark process. Finally, conclusions are made.

2. Random forests

Random forests (RF) is a tree-structured ensemble learning method, which was first proposed by Breiman in 2001. Since then, it has been widely used for both classification and regression in various areas, due to its superior performance and simple structure. Two random selection procedures in random forests make it perform well compared with the other methods. One is bootstrap aggregation, also known as bagging, the other one is random selection of features. Bagging, proposed by Breiman in 1994 [26], is designed to improve the stability and accuracy of machine learning algorithms. The main idea of bagging is generating a number of training sets by drawing random samples with replacement (bootstrap) from the original training set. Then various models are trained by using the above bootstrap samples and the results of the models are aggregated to make the final decision. The classic application of bagging is random forests, which combines bagging with decision tree method. But the training procedure of decision tree in random forests is modified, with random selection of features at each node of the tree and without pruning. Due to the two random characteristics, random forests is able to reduce the variance and simultaneously maintain the low bias.

The detailed steps of random forests are shown as follows:

(1) Generate a set of new training sets by randomly sample with replacement (bootstrap) from the original training set.
(2) For each new training set, constructed a tree with random selection of features at each node of the tree and without pruning.
(3) After a large number of trees are generated, the new data is predicted by aggregating the results of all the trees, with majority voting strategy.

A typical structure of random forests is shown in Fig. 1.