Pattern Recognition Letters 80 (2016) 179–187

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Fast structural ensemble for One-Class Classification☆

Jiachen Liu[a], Qiguang Miao[a,*], Yanan Sun[a], Jianfeng Song[a], Yining Quan[a]

[a] School of Computer Science and Technology, Xidian University, 2nd Taibai South Road, Xi'an 710071, China

ARTICLE INFO

Article history:
Available online 9 July 2016

Keywords:
One-class classifier
Clustering
Structural ensemble
Divide-and-conquer

ABSTRACT

One of the most important issues of One-Class Classification (OCC) algorithm is how to capture the characteristics of the positive class. Existing structural or clustering based ensemble OCC algorithms build description models for every cluster of the training dataset. However, the introduction of clustering algorithm also causes some problems, such as the determination of the number of clusters and the additional computational complexity. In this paper, we propose Fast Structural Ensemble One-Class Classifier (FS-EOCC) which is a fast framework for converting a common OCC algorithm to structural ensemble OCC algorithm. FS-EOCC adopts two rounds of complementary clustering with fixed number of clusters. This number is calculated according to the number of training samples and the complexity of the base OCC algorithm. Each partition found in the previous step is used to train one base OCC model. Finally all base models are modularly aggregated to build the structural OCC model. Experimental results show that FS-EOCC outperforms existing structural or clustering based OCC algorithms and state-of-the-art non-structural OCC algorithms. The comparison of running time for these algorithms indicates that FS-EOCC is an efficient framework because the cost of converting a common OCC algorithm to a structural OCC algorithm is small and acceptable.

1. Introduction

One-Class Classification (OCC) is a machine learning task which aims at distinguish samples of positive class from samples of negative class [1,2]. OCC is a tough problem because negative samples are often severely under-sampled or even totally absent. However OCC techniques are useful in practice when negative samples are hard to be obtained. So far, there are many OCC algorithms available such as kernel based One-Class Support Vector Machine (OC-SVM) [3] and Support Vector Data Descriptor (SVDD) [4], density based Gaussian estimator and Mixture of Gaussian (MoG) method etc.. Recently, many researchers are concentrating on mining and describing the inner characteristics of the positive class to build more robust OCC models. Examples include diversity [5], structural information [6] and multi-modality [7] contained in positive samples.

In this paper, we focus on the structural or clustering based OCC algorithms. These kinds of OCC algorithms are applied in real world problems [8–10] for its abilities of describing data with multi-modality and multi-density. However, structural OCC algorithms often have higher computational complexities due to the clustering step integrated in them. Moreover, the determination of the cluster number in the training dataset is not trivial. Existing structural OCC algorithms often use hierarchical or repeatedly clustering to determine the number of clusters. These methods are time-consuming and their accuracies could not be guaranteed [6].

To address the above problems of structural OCC algorithms, we develop a fast and parameter-free framework to construct a structural OCC model. The proposed method is named Fast Structural Ensemble One-Class Classifier (FS-EOCC). There are three key aspects and also main contributions of FS-EOCC:

- The number of clusters is calculated first instead of the difficult and time consuming step of determining it. The number of clusters used in FS-EOCC depends on the size of the training dataset and the computational complexity of the base OCC algorithm.
- FS-EOCC adopts two rounds of complementary clustering algorithms with a small MST (Minimum Spanning Tree) structure to better describe the positive class. And the gaps between base OCC models could be avoided.
- The modular ensemble scheme is used in the aggregation step of FS-EOCC to build the complete description model, which is similar to existing structural OCC algorithms. But FS-EOCC is much faster than them because it keeps the clustering step as simple as possible.

The rest of this paper is organized as follows. In Section 2, related OCC algorithms are reviewed. Then FS-EOCC is proposed

☆ This paper has been recommended for acceptance by Dr. G. Moser.

* Corresponding author.
E-mail address: qgmiao@xidian.edu.cn, qgmiao@126.com (Q. Miao).

http://dx.doi.org/10.1016/j.patrec.2016.06.028
0167-8655/

in Section 3. Section 4 gives the experimental results of FS-EOCC comparing with state-of-art structural or ensemble OCC algorithms. Finally we conclude the paper in Section 5.

2. Related works

The first OCC algorithm was proposed in [11], and then several kinds of OCC algorithms were developed in the past two decades. Some of the most popular ones include: density based Parzen window estimator [12], Gaussian estimator [13] and its extension Mixture-of-Gaussian (MoG) [14]; boundary based OC-SVM [3], SVDD [4] and the nearest neighbor method [1]; reconstruction based k-means method [1] etc.. In this paper we focus on structural OCC algorithms, details about common OCC algorithms could be found in review literature [1,2]. OCC algorithms have been applied in many practical problems e.g. image classification [15], intrusion detection [16], malware detection [17] and remote sensing [18].

Along with the rapid development of ensemble learning, the ensemble of OCC is becoming an important research field. Some recent ensemble OCC algorithms e.g. the combination of OCC and random forest method [19] are trying to introduce popular ensemble learning algorithms to OCC methods. Advanced ensemble learning techniques such as ensemble pruning [20,21] and diversity [22] are also introduced to refining OCC model. These algorithms have also been applied in practical problems e.g. medical image analysis [23]. Moreover, ensembles of OCC models could solve the multi-class classification problems with more flexible options such as rejecting and robust to noise [24].

Among so many ensemble OCC algorithms above, there is a kind of ensemble methods called structural ensemble. The structural ensemble of OCC model utilizes several base OCC model to describe the positive class by groups. This kind of ensemble OCC model could better capture complex characteristics of the training samples. Wang et al. [25] proposed the first structured one-class classifier (TOCC) which is one of the earliest attempts to consider the data structures in OCC and use multiple base OCC models to describe training samples. TOCC first finds the clusters in the positive class using agglomerative hierarchical clustering (AHC). And then multiple ellipsoids corresponding to the clusters are optimized by solving a series of second-order cone programming problems. TOCC used hierarchical clustering to avoid repeated running of clustering algorithms, however the computing complexity of hierarchical clustering algorithms are higher than flat ones [27]. Moreover, TOCC chooses the cluster number by cutting the dendrogram at the knee point. This method has some limitations e.g. it cannot determine if a dataset contains only a single cluster [6].

Krawczyk et al. [6] proposed one-class clustering-based ensemble (OCClustE) method. OCClustE is a framework of OCC based on flat clustering algorithms. The clustering algorithm, base classifier and fusion method of OCClustE can be chosen by the users. OCClustE chooses the number of clusters via comparing entropies of the membership values. However as the authors indicated in their paper: "*We acknowledge that the entropy criterion is not a perfect solution for estimating the number of clusters.*", the estimation of the number of clusters significantly affects the performances of clustering-based OCC. OCClustE could use kernel fuzzy clustering and fuzzy membership values for weight calculation of the base OCC models to enhance its performance. Recently, the authors of OCClustE discussed this problem in depth and compared many clustering model selection methods to enhance the performance of OCClustE [26]. Similar approaches such as [9,10] applied clustering based OCC algorithms to information retrieval and compliance verification.

Another important OCC algorithm based on the mining of structural information contained in the positive class is Minimal Spanning Tree based Class Descriptor (MST_CD) [28]. MST_CD makes use of the Minimal Spanning Tree (MST) of all the positive samples. The predictions of testing samples depend on their distances to the closest edges of the MST. MST_CD could capture the characteristics of the positive class especially in high-dimensional space or the sample size is small. The main limitation of MST_CD is that MST is not flexible enough to be used to describe complex distributions of the positive class. For example, some false positives are hard to avoid if there are multiple clusters in the positive class. Moreover, MST_CD is not robust enough to handle noise samples because all samples are included in MST [29]. MST_CD motivates some recent structural ensemble OCC algorithms. Liu *et al.* [30] proposed a modular ensemble OCC algorithm based on density analysis and MST structures. The MST structures are built with modified distance measures considering local density so the dividing of training samples is more flexible.

3. Fast structural ensemble for one-class classification

3.1. Overview of FS-EOCC

The main limitations of existing clustering based or structural OCC are two folds:

(1) The existing structural OCC algorithms are of high computational complexities. For example, OCClustE invokes the clustering algorithm repeatedly with different number of clusters and choose one of them via comparing entropy values of different clustering results. TOCC adopts hierarchical clustering and knee-point method to determine the number of clusters. Although for TOCC, the hierarchical clustering runs only once but it is a clustering algorithm with high complexity, because it finds the complete dendrogram rather than the final clustering result.
(2) The number of clusters must be determined, which is hard or even not reasonable under some situations. On one hand, the determination of the number of clusters is an unsolved problem [31] and then errors introduced by clustering will have negative impacts on the OCC models. On the other hand, the clustering step of structural OCC is not reasonable if the distribution of the positive class does not have multiple natural clusters.

To tackle these problems above, we construct FS-EOCC as the following steps:

(1) **First-round partitioning.** The clustering algorithm is performed on the training dataset with pre-calculated number of clusters. The calculation of the number of clusters is based on the computational complexity minimization and the maximum reasonable cluster number.
(2) **Preparing of second-round partitioning.** A minimum spanning tree is constructed based on the centroids of the first-round clustering. The initial centroids of the second-round clustering are the midpoints of all edges of the minimum spanning tree.
(3) **Second-round partitioning.** The clustering algorithm is performed again on the training dataset with the initial centroids which are calculated in step (2). The aim of the second-round partitioning is to describe the gaps of the describing boundary if constructed with only the result of the first-round partitioning.
(4) **Building and combining of base models.** Each partition of the training dataset found in the two rounds of clustering is used to build a base OCC model. For each base OCC model, the positive training class consists of samples of one partition and all the other samples are used as negative samples. The description boundaries of trained base models are