Neurocomputing 171 (2016) 262–276

Contents lists available at ScienceDirect

Neurocomputing

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

Modular ensembles for one-class classification based on density analysis

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

School of Computer Science and Technology, Xidian University, Xi'an 710071, PR China

ARTICLE INFO

Article history:
Received 16 January 2015
Received in revised form
13 May 2015
Accepted 21 June 2015
Communicated by: Shiguang Shan
Available online 2 July 2015

Keywords:
One-class classifier
Modular ensemble
Density analysis

ABSTRACT

One-Class Classification (OCC) is an important machine learning task. It studies a special classification problem that training samples from only one class, named target class, are available or reliable. Recently, various OCC algorithms have been proposed, however many of them do not adequately deal with multi-modality, multi-density, the noise and arbitrarily shaped distributions of the target class. In this paper, we propose a novel Density Based Modular Ensemble One-class Classifier (DBM-EOC) algorithm which is motivated by density analysis, divide-and-conquer method and ensemble learning. DBM-EOC first performs density analysis on training samples to obtain a minimal spanning tree using density characteristics of the target class. On this basis, DBM-EOC automatically identifies clusters, multi-density distributions and the noise in training samples using extreme value analysis. Then target samples are categorized into several groups called Local Dense Subset (LDS). Samples in each LDS are close to each other and their local densities are similar. A simple base OCC model e.g. the Gaussian estimator is built for each LDS afterwards. Finally all the base classifiers are modularly aggregated to construct the DBM-EOC model. We experimentally evaluate DBM-EOC with 6 state-of-art OCC algorithms on 5 synthetic datasets, 18 UCI benchmark datasets and the MNIST dataset. The results show that DBM-EOC outperforms other competitors in majority cases especially when the datasets are multi-modality, multi-density or noisy.

1. Introduction

Machine learning is one of the most attractive research topics in computer science. There are many different learning schemes studied in machine learning field. Lots of successful applications have been developed according to them. An important and interesting learning problem is called One-Class Classification (OCC) [1] which is similar to anomaly detection [2] or novelty detection [3]. However, anomaly detection and novelty detection are totally regardless of labels, but depend on pre-defined constraints on abnormality. With comparison, OCC is more like a supervised learning approach because all training data are labeled as "target class", whose objective is to classify target samples from samples which are not belonging to the target class, named "outlier class". Typical OCC tasks do not have any outlier samples available during the training phase. As a consequence, OCC is a difficult machine learning problem because it has to build a precisely descriptive instead of discriminant model of the target class with enough generalization ability. OCC has been widely applied to solve many practical one-class problems, such as intrusion detection with only normal network traffic samples available [4] and jet engine faults detection with only data known to be normal [5]. OCC can also be used in binary classification tasks in which samples from one class is too few or severely under-sampled, or the so-called "open-set recognition" in which not all classes are known at the training time. Examples include imbalanced classification [6,7], speaker verification [8] and open-set image recognition [9]. For multi-class classification, an OCC model could be trained for each class and outputs of these based models are aggregated to a final classification result [10,11].

In essence, OCC is a description task rather than classification, therefore OCC models should have abilities to describe the peculiarities of target distribution such as multi-model, multi-density and so on. However, many existing OCC algorithms could not handle such situations very well because of a lack of sufficient description power. Here we use a two-dimensional synthetic dataset shown in Fig. 1 to illustrate this problem.

The synthetic dataset used here is named "four bananas" which consists of two pairs of banana-shaped distributions for the target class and 333 uniformly distributed outliers. Each banana-shaped target distribution contains 250 samples. The smaller pair is denser than the larger pair. Thus, the "four bananas" dataset has characteristics of multi-model, multi-density and arbitrary shapes. To illustrate their descriptive power, three state-of-art OCC

* Corresponding author. Tel.: +86 18681880036.
E-mail address: qgmiao@xidian.edu.cn (Q. Miao).

http://dx.doi.org/10.1016/j.neucom.2015.06.037
0925-2312/

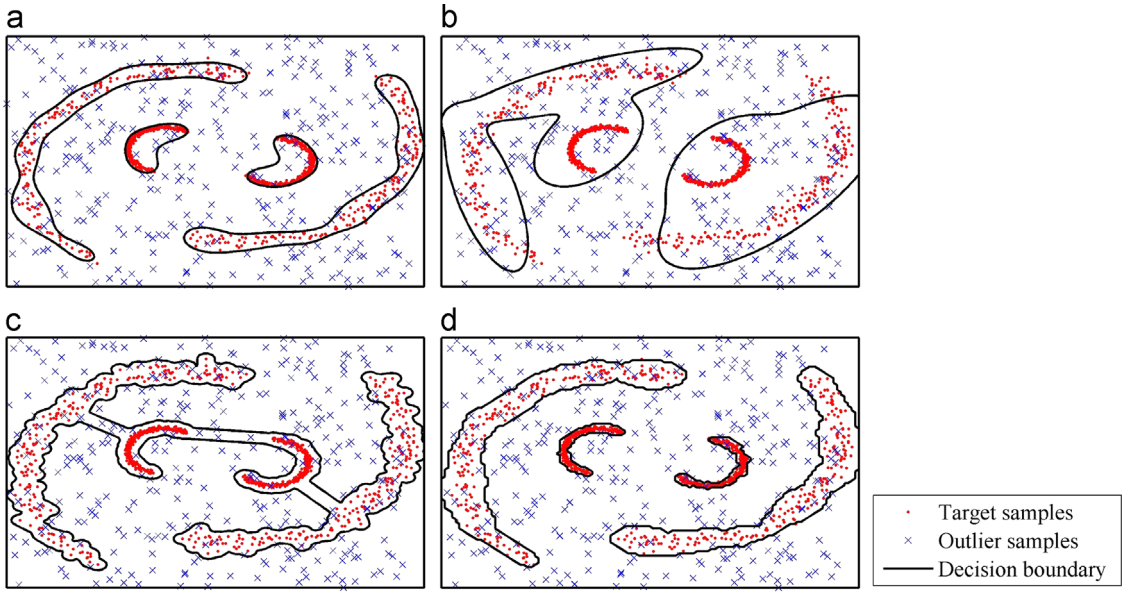

Fig. 1. Illustrations of different OCC algorithms on the "four bananas" dataset, (a) OC-SVM, (b) MoG, (c) MST_CD, and (d) DBM-EOC.

algorithms are selected including popular One-Class Support Vector Machine (OC-SVM) [12], the typical multi-modality Mixture-of-Gaussian (MoG) algorithm and the tree-structured Minimal Spanning Tree based Class Descriptor (MST_CD) [13] shown in Fig. 1(a)–(c). The Radial Basis Function (RBF) kernel and the default kernel width of LibSVM [14] are used for OC-SVM and MoG is trained using a Expection Maximization (EM) method. Decision boundaries of the trained OCC models are shown in Fig. 1 with solid lines. From Fig. 1(a) we can see that the decision boundary of OC-SVM model is too loose for the smaller pair of banana-shaped target distributions. This is because target distributions with different density levels need different kernel widths for RBF kernel of OC-SVM, however OC-SVM has only one global kernel function. The same problem exists as well in MST_CD as shown in Fig. 1(c), because the decision threshold is a global value. The descriptive model given by MST_CD has some additional false positive regions between two clusters, because there are edges of the minimal spanning tree. From Fig. 1(b) we can see that the MoG model fails to detect the right clusters and the arbitrarily shaped distribution of the target class even if a proper number of clusters is given to it.

Although the synthetic data in 2-dimensional space do not exactly represent the situations of practical high-dimensional datasets. We are trying to emphasize the importance of data model of the target class for OCC algorithms. As a recent book on outlier analysis [15] outlines: "*The data model is everything. … Clearly, the choice of the data model is crucial. An incorrect choice of data model may lead to poor results.*" As aforementioned, outlier analysis is very similar to OCC but the main difference is the absence of labels in outlier analysis. The conclusions for outlier analysis are also applicable for OCC, and it is undoubtable that OCC algorithms would not perform well if the characteristics of multi-modality, multi-density, etc. are not considered in the model of target class.

To better deal with multi-modality, multi-density, arbitrarily shaped distributions and the noise of target samples, we propose an OCC algorithm named Density Based Modular Ensemble One-class Classifier (DBM-EOC) as shown in Fig. 1(d). DBM-EOC first performs density analysis on the target class to build a tree-shaped structure to represent the density distribution of the target samples. On this basis, several Local Dense Subsets (LDS) of the target class are constructed to describe groups of target samples which are close to each other and have similar density. The clusters, the noise and distributions with different density levels are automatically detected at a time. Then base OCC models are trained for each LDS by divide-and-conquer method. Finally all the base models are modularly aggregated to construct the DBM-EOC model. The main contributions of our proposed DBM-EOC are listed as follows:

- DBM-EOC detects multi-modality, multi-density and arbitrarily shaped distribution of training samples from the tree-structure obtained by density analysis.
- DBM-EOC automatically removes the noise in the process of constructing LDS. So DBM-EOC is more robust to the noise in target samples which may cause significant false positives. Most existing OCC algorithms remove the noise via a pre-defined and fixed target rejection ratio, while DBM-EOC uses one-dimensional extreme value analysis which is more flexible.
- DBM-EOC aggregates several simple base OCC models trained on every LDS by modular ensemble. The OCC problem is solved via a divide-and-conquer way, which leads to lower computational complexity and the ability of handling arbitrarily shaped target distributions.

The paper is organized as follows. A review of existing OCC algorithms is given in Section 2. In Section 3, the proposed DBM-EOC is described in details, including the discussions of our motivations in Section 3.1, the density analysis in Section 3.2, the modular ensemble in Section 3.3 and the analysis of computational complexity in Section 3.4. The experimental protocol and results are presented in Section 4. Finally Section 5 concludes the paper.

2. Related works

So far, researchers have proposed various OCC algorithms, which have been tested to be useful in many practical applications. Detailed descriptions of OCC could be found in some review articles [1–3,16]. OCC is similar to anomaly detection, novelty detection and outlier detection in the literatures of machine learning, data mining and pattern recognition. In some research, terminologies OCC, anomaly detection, novelty detection and