Expert Systems with Applications 42 (2015) 4840–4850

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Two approaches for novelty detection using random forest

Qi-Feng Zhou [a], Hao Zhou [a], Yong-Peng Ning [a], Fan Yang [a,*], Tao Li [b,*]

[a] Department of Automation, Xiamen University, Xiamen 361005, China
[b] School of Computer Science, Florida International University, Miami, FL 33199, USA

ARTICLE INFO

Article history:
Available online 19 February 2015

Keywords:
Novelty detection
Ensemble learning
Random forest
Proximity matrix

ABSTRACT

In many online classification tasks or non-exhaustive learning, it is often impossible to define a training set with a complete set of classes. The presence of new classes as well as the novelties caused by data errors can severely affect the performance of classifiers. Traditional proximity-based approaches usually utilize the distance to measure the proximity of different samples. In this study, we propose a framework that uses ensemble learning to detect novelty based on Random Forest (RF). The proposed framework is based on the observation that an ensemble of classifiers can provide a kind of metric to characterize different classes and measure their proximity. In particular, we apply ensemble methods with the decision tree as base classifiers and present two specific approaches, RFV and RFP, based on random forest. RFV uses the vote distribution of RF on a testing sample, and RFP takes the proximity matrix of RF as a special kernel metric to discover the novelty. The proposed approaches are compared against two common approaches: support vector domain description (SVDD) and Gaussian Mixed Model (GMM) on one artificial data set and five benchmark data sets. The experimental results show that the proposed methods achieve better performance in terms of accuracy and recall.

1. Introduction

Classification as a fundamental task in machine learning is a pervasive problem that encompasses many different applications such as image analysis, character recognition, disease diagnosis, and human identification (Dietterich, 1997). In general, the mathematical formulation of classification can be described as follows. Given m points (x_i, y_i), $i \in M = \{1, 2, \ldots, m\}$, in the $(n+1)$-dimensional real space R^{n+1}, where x_i is a n-dimensional vector and y_i a scalar. Each point can be interpreted as a sample; the coordinates of the vector x_i are the values of the attributes; and the target y_i denotes the class to which the sample belongs, $y_i \in Y = \{Y_1, Y_2, \ldots, Y_n\}$ where Y is the class label set. Classification aims to learn a target function that maps each attribute set x_i to one of the predefined class label y_i. The target function is also known as a classification model that is used to predict the class label of unknown samples.

A classification task consists of two stages: training and testing. In the training stage, a training set containing samples whose labels are known must be provided to build the classification model. In the testing stage, the learned model is applied to the test set containing unknown samples. A common assumption is that the training set and the test set are drawn from the same distribution and share the same class label set so that the classification model could have a good generalization capability.

However, there are some cases that make the aforementioned assumption invalid: (1) abnormal data are often mixed with normal data because of data measurement and collection error; (2) new classes different from the trained classes emerge in the test stage. Both cases will generate anomalies or novelties, although they differ in that the former case usually generates novelties in smaller size and possible emerging in the training stage. For novelties emerging in the training stage, novelty detection (and removal) becomes actually a part of data preprocessing. The appearance of new classes is often more critical, as it could change the data distribution and degrade the classification performance. For instance, in network intrusion, ordinary attacks usually have the purpose of disabling computers and they are often sampled as the training data to build the detection model. In the testing stage, however, a number of new types of attacks may arise such as those attacks aiming to steal users' information. These new attacks will be consistently misclassified unless they are labeled as new classes and the detection model undergoes re-training. In

* Corresponding authors.
E-mail addresses: zhouqf@xmu.edu.cn (Q.-F. Zhou), colinzhou2013@gmail.com (H. Zhou), 961331014@qq.com (Y.-P. Ning), yang@xmu.edu.cn (F. Yang), taoli@cs.fiu.edu (T. Li).

http://dx.doi.org/10.1016/j.eswa.2014.12.028
0957-4174/

some cases, the new classes may replace some existing classes, making the learned classifier completely invalid. Similar examples can be found in fraud detection, ecosystem disturbance, public health, and medicine (Pang-Ning, Steinbach, & Kumar, 2006; Yexi Jiang & Jian Xu, 2014). In these applications, the novelties often contain valuable information and are the focus to which we need pay more attention. Therefore, novelty detection is of great importance and is one of the fundamental requirements of a good classification or identification system.

In this paper, we propose a framework to detect novelty by means of ensemble learning and present two specific approaches based on random forest. Our work focuses on the novelties from new classes but the approach is applicable to all types of novelties. The rest of the paper is structured as follows. Section 2 briefly reviews the related work. Sections 3 and 4 describe ensemble learning and random forest that form our basis for novelty detection. In Section 5, two specific approaches using random forest, which are named RFV and RFP, respectively, are presented with details. Section 6 presents the experiments and related analysis. Conclusions are drawn in Section 7.

2. Related work

Although novelty detection has a long history, most of the approaches are originally proposed as a part of data preprocessing to detect anomalous objects. Clearly, under these circumstances, novelty detection is not necessarily linked to classification but is used as a general technique for data preparation. Existing approaches mainly consist of the following categories: statistical, proximity-based, density-based, and clustering-based approaches (Pang-Ning et al., 2006); where each category has its strengths and weaknesses. However, not all the approaches can be used for classification tasks; e.g., clustering-based approaches are usually geared towards unsupervised learning, instead of classification.

In statistics, novelty detection is known as outlier detection as novelties usually take the form of outliers. The article by Beckman and Cook (1983) provides a general overview of how statisticians view the subject of outlier detection. An extensive survey about outlier detection methods is provided by Hodge and Austin (2004). The statistical approaches are generally based on building a probability distribution model and considering how well the sample fits the model. The model can be very effective when an appropriate distribution is chosen. In practice, prior knowledge is often required for accurate estimation. One of the frequently used models is Gaussian Mixed Model (GMM) (Lauer, 2001), which assumes the objects following a mixture of Gaussian distributions. Those samples with lower probabilities than a specific threshold are viewed as novelties. Unfortunately, in reality, there are many datasets with non-standard distributions, and applying statistical models on them can perform poorly.

The proximity- and density-based approaches have similar mechanisms, searching for outliers that are distant from most objects. Because these approaches are essentially classification algorithms, it is easy to apply them in classification. One simple practice is to use the k-nearest neighbor (Ramaswamy, Rastogi, & Shim, 2000) classifier, which can easily identify whether a test sample is a novelty or not. One limitation of these categories of the approaches is that they require a measure to calculate the proximity between samples. For example, a single decision tree is inappropriate for novelty detection, as the proximity between the samples is difficult to measure.

Recently a new category of approaches to novelty detection has been developing rapidly. These approaches are characterized by using a variety of neural networks, including multi-layer perceptrons, self-organizing maps, radial basis function networks and support vector machines (SVM) (Fiore, Palmieri, Castiglione, & De Santis, 2013; Markou & Singh, 2003). Among these neural networks, support vector domain description (SVDD) (Schölkopf, Williamson, Smola, Shawe-Taylor, & Platt, 1999; Spinosa & Carvalho, 2005) is particularly suitable for novelty detection. SVDD is derived from one-class research (Chen, Zhou, & Huang, 2001; Kemmler, Rodner, Wacker, & Denzler, 2013; Manevitz & Yousef, 2002) and uses a hypersphere to enclose all objects in one target class with a minimal volume by minimizing the structural risk. When a sample is found to be outside of the target hypersphere, it is identified as a novelty. Some neural networks based approaches also utilize ensemble learning to improve the performance of novelty detection (Markou & Singh, 2003). Specifically, a collection of different neural networks, instead of a single one, are trained as the classifiers and a novelty is a sample that has low outputs on all classifiers. Neural networks based approaches are hindered by the computational complexity and the volume of the network considered to achieve best performance.

It is noticed that an ensemble of classifiers can actually provide a kind of metric to measure the proximity between samples (or classes). Assuming that each classifier casts a vote on a sample, the votes of all the classifiers may have different distributions with respect to the samples from different classes. Intuitively, when the majority of the classifiers cast the same vote for one class, the prediction for the sample tends to be of high confidence. In that case, the test sample with a high probability belongs to a known class. By contrast, when there is a great divergence in the votes, the probability for the sample belonging to any known class is low, and then it is likely to be a novelty. This shows that the difference of the votes on samples can measure their proximity. As in the traditional proximity-based approaches, the vote-based proximity can also be used to detect potential novelties. More importantly, this kind of proximity metric is independent of the specific classifiers; as a result, some classification algorithms, which do not work based on the proximity or distance, can be used to detect novelty. An illustrative case is an ensemble of trees. As discussed before, a single tree as a classifier is not appropriate for novelty detection, but an ensemble of trees may be feasible. In next section, we will systematically discuss applying ensemble learning to novelty detection and analyze the feasibility.

3. Ensemble learning and confidence

Ensemble learning (Maclin & Opitz, 2011; Meina et al., 2013; Polikar, 2006; Rokach, 2010) integrates multiple models to obtain better performance than that could be obtained from any of the constituent model. In classification, ensemble methods construct a set of base classifiers from training data and perform classification by taking a vote on the predictions made by each base classifier. Formally

$$C(x) = Vote(C_1(x), C_2(x), \ldots, C_k(x)) \tag{1}$$

where $C_i(x)$ is the prediction made by the ith base classifier. A test sample x is classified by taking a majority vote on the individual predictions or by weighting each prediction with the accuracy of the base classifier. Through bias-variance decomposition (Friedman, 1997), it could explain why ensemble methods tend to perform better than any single model. Common techniques to construct an ensemble system include Bagging (Breiman, 1996), Boosting, and AdaBoost (Cortés, Martínez, & Rubio, 2013). An eminent application of ensemble learning is random forest.