Neurocomputing 201 (2016) 12–18

Contents lists available at ScienceDirect

Neurocomputing

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

Probabilistic framework of visual anomaly detection for unbalanced data ☆

Yongxiong Wang *, Xuan Li, Xueming Ding

Key Laboratory of Modern Optical System, and Engineering Research Centre of Optical Instrument and System, Ministry of Education, University of Shanghai for Science and Technology, 200093 Shanghai, China

ARTICLE INFO

Article history:
Received 30 December 2014
Received in revised form
10 October 2015
Accepted 28 March 2016
Communicated by Xu Zhao
Available online 4 May 2016

Keywords:
Anomaly detection
Posterior probability
Parameter learning
Unbalance distribution

ABSTRACT

This paper proposes a novel probabilistic detection framework of weighted combining semi-supervised *k*-means clustering and Posterior Probability SVM (PPSVM) for unbalanced data based on robot vision. Within the framework, an algorithm for learning synchronously the *k* in *k*-means and features is introduced based on hybrid wrapper and filter criterion. Then the optimal hierarchical probabilistic model by combining *k*-means and PPSVM is used to anomaly detection so as to alleviate the problems of imbalanced data with small samples, improve the detection accuracy, and deal with the difficult problem of defining the anomaly classes. The other contributions of our approach include the following three aspects: (1) it classifies anomaly candidates by using their class probability distributions rather than the direct extracted features; (2) the relevant classes are automatically built by learning the samples' multimodal Gaussian distribution; and (3) the cost-sensitive idea and filter criterion are integrated in learning *k* and features via cost function of Tabu search. Experimental results on real-world data sets show the proposed approach obtains a satisfactory detection performance within limited time in inspecting the condition of Heating, and Ventilation and Air-Conditioning (HVAC) ductwork.

1. Introduction

With the wide application of computer vision, the anomaly detection via vision or image has a wide range of applications such as sewer detection [1], image analysis [2], surface defect detection [3], video surveillance [4], and industrial damage detection [3]. The most common algorithms of anomaly detection are based on machine learning techniques which have extensively been studied and used because of its low false positive rates and other advantages [2,5]. However, anomaly detection based on robot vision has limitations. Firstly, defining a normal region which contains all possible normal behaviors is maybe very difficult in actual anomaly detection [5], especially in visual anomaly detection. Secondly, most of anomaly detection or classification problems include class imbalance and are naturally cost-sensitive. This situation called within-class and between-class imbalanced data distributions should deteriorate detection performance in difficult-to-classify problems with overlapping and/or in the absence of a sufficient number of minority training samples [6,7]. Generally, the anomaly samples which need to be recognized maybe be wrongly classified to the normal samples without reducing the total recognition rate. At last, the diversities of feature distributions (or multimodal Gaussian distribution) within the anomaly or normal class are difficult to be precisely defined artificially and lead to the within-class imbalance [8]. This is also one of the main reasons to lead to poor detection performance.

When it is difficult to fine the criteria for normal and anomaly class and the knowledge of anomaly classes is incomplete, the clustering algorithms like *k*-means, Self-Organizing Map (SOM), fuzzy *C*-mean, and so on [5,9] are usually employed to define the anomaly groups automatically. The *k*-means can guarantee at least a local minimum of the criterion function in anomaly detection. And it can ensure each training sample is associated with only one cluster and need not artificially label samples. Other researcher focus on the difficult scenarios where discovery and classification of rare class need to be tackled jointly [10,11]. Haines and Xiang employed an active learning method for classifying and discovering unknown rare classes. Our aim is to jointly select the feature sets of compute vision and classify rare class which includes multi-subclasses.

We use *k*-means to preprocess the feature distributions of samples and acquire the initial decision boundary of the anomaly sub-classes. However, the unbalanced data distributions in anomaly inspection will arise two problems in *k*-means clustering [9]: (1) the class unbalance (dominance) problem when the

☆This work is supported in part by the National Natural Science Foundation of China under Grant nos 61374039, 61403254 and Hujiang Foundation of China (C14002, B1402/D1402 and D15009).

* Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2016.03.038
0925-2312/

training data have a large number of samples from one particular class (majority class) and very few samples from the remaining classes (minority class), and (2) the forced assignment problem when k parameter in k-means is set to a value that is considerably less than the inherent number of natural groupings within the training data. Some overlapping minority groups within a cluster may not be captured and the samples from the minority classes are probably misclassified. In [9], cascading k-means clustering and the ID3 decision tree is used to relief the two problems and acquire high performance in anomaly detection. Additionally, the appropriate k in k-means is very important to improve the detection accuracy or reduce the false positive rate in anomaly detection.

There are previously several algorithms to determine k in k-means automatically. G-means algorithm is applied to learn the k in k-means based on a statistical test of fitting a Gaussian distribution [12]. Bischof et al. [13] use a minimum description length framework to choose k, where their algorithm starts with a large value for k and removes centers (reduces k) whenever that choice reduces the description length.

A number of approaches have been proposed to relief imbalanced data: (1) Data-level approaches [14], such as oversampling the small class or undersampling the large class. (2) Cost-sensitive approaches [15], one way is to use different parameters C in the cost function for the two classes in classifiers. (3) Fusion and cascading of multiple machine learning methods [9], which have a better performance yield over individual methods. (4) The feature selection for imbalanced data sets [16]. Especially, it is beneficial to handling the moderately high-dimensional imbalanced data sets with small samples [16].

Tabu Search (TS) would be a promising technique for feature selection in order to obtain the discriminating feature subsets and improve computation efficiency [17]. It has shown that TS not only could obtain the optimal or near-optimal solution, but also less require the computational effort than other sub-optimal and genetic algorithm based search methods. However, [17] did not take into account the problem of unbalanced data distribution.

In this paper we present a learning probabilistic diagnosis framework which combine semi-supervised k-means clustering and posterior probability SVM for unbalanced data based on robot vision. Within the framework, we select a suitable k of k-means and the appropriate features of image based on TS, which has a cost function associated with maximization of between-class distance and cost-sensitive minimum misclassification. To improve the real-time, we use TS to learn discriminating features for great reducing the time of image process and apply a hierarchical strategy to rapidly discard the most normal samples under predetermined threshold using probability k-means detection method in the first level. And then the decision boundary to classify the anomaly samples is refined by using supervised PPSVM in anomaly detection.

The advantages of our algorithm are that we can synchronously learn k of k-means and feature based on hybrid wrapper and filter criterion. In anomaly detection of unbalanced data with multi-subclasses, we can improve the classification accuracy of minority classes (anomaly) at keeping the accuracy of majority class, and obtain a meaningful anomaly score for each test instance.

The rest of this paper is organized as follows. Section 2 describes the proposed systematic framework which includes brief introduction of the k-means, PPSVM learning-based anomaly detection methods, and hierarchical detection algorithm. In Section 3, we present the learning method of k and feature selection by the TS. Experimental setup and results are presented in Section 4. Section 5 concludes the work.

2. The proposed probability detection method for unbalanced data via robot vision

Firstly, we concisely discuss semi-supervised k-means clustering [18] and PPSVM [19] for anomaly detection in the section. Then we propose a hierarchical anomaly detection approach based on the two methods.

2.1. Anomaly detection with semi-supervised k-means feature-clustering

In many cases, the knowledge of anomaly classes is incomplete. The semi-supervised clustering by seeding [18] can group data using labeled data to generate seed clusters that initialize a clustering algorithm. In our k-means method, the seed cluster is only used for the initialization. The steps in the semi-supervised k-means-based anomaly detection are as follows:

The semi-supervised *k*-means-based anomaly detection

1: Initialize k cluster centers, r_i, using the mean of ith in seed set according to supervision, for $i=1,\ldots, k$. (We assume there is at least one seed point that belongs to each class.)
2: Assign cluster: Assign each data point x to the nearest cluster h, for $i=\arg\min_i \|x-r_i\|^2$.
3: Update each cluster center r_i as the mean of all date that belongs to it.
4: Repeat step 2–4 until cluster centers are stable.
5: For each test sample Z
a. Compute the Euclidean distance $D(r_i,Z)=\|r_i-Z\|^2$, $i=1,\ldots, l$, find the cluster r_i, that is the closest to Z.
b. Classify Z using either the threshold rule or the Bayes decision rule.

The threshold rule: Assign $Z\rightarrow 1$ (Z belongs to the cluster r_i) if $P(\omega_i\in 1\,|\,Z\in C_i)>\tau$, and τ is a threshold; Otherwise $Z\rightarrow 0$, where "0" and "1" represent normal and anomaly class, respectively. $\omega_i\in 1$ represents the anomaly class in the cluster r_i, $P(\omega_i\in 1\,|\,Z\in C_i)$ represents the probability of anomaly samples in r_i.

The Bayes decision rule: Assign $Z\rightarrow 1$ (Z belongs to the cluster r_i) if $P(\omega_i\in 1\,|\,Z\in C_i)>P(\omega_i\in 0\,|\,Z\in C_i)$; Otherwise $Z\rightarrow 0$, where ω_i represents the anomaly class in the cluster r_i, $P(\omega_i\in 0\,|\,Z\in C_i)$ represents the probability of normal samples in r_i.

2.2. Anomaly detection with posterior probability method of SVM

The advantage of a PPSVM is the fact that it is closer to the Bayes optimal without knowing the distributions [20,21] and it can be used to solve classification problems where the classes and samples are not equally distributed [21,22]. Instead of predicting the label, we used PPSVM to give the approximating posterior class probability $P^{svm}(\omega_i=1\,|\,x)$ by a sigmoid function [21].

$$P^{svm}(\omega_i=1\,|\,x)\approx P_{A,B}(f)\equiv\frac{1}{1+\exp(Af+B)},\ \text{where } f=f(x). \tag{1}$$

Let each f_i be an estimate of $f(x_i)$. The best parameter $\theta^*=(A^*,B^*)$ is acquired by solving the following maximum likelihood problem with regularize term (with N_+ of the θ_i's positive, and N_- negative):

$$\min_{\theta=(A,B)} F(\theta)=-\sum_{i=1}^{l}(t_i\log(p_i)+(1-t_i)\log(1-p_i)), \tag{2}$$