Neurocomputing 133 (2014) 483–490

Contents lists available at ScienceDirect

Neurocomputing

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

Regularized correntropy criterion based feature extraction for novelty detection

Hong-Jie Xing*, Huan-Ru Ren

Key Laboratory of Machine Learning and Computational Intelligence, College of Mathematics and Computer Science, Hebei University, Baoding 071002, China

ARTICLE INFO

Article history:
Received 8 November 2012
Received in revised form
4 December 2013
Accepted 7 December 2013
Communicated by Ran He
Available online 10 January 2014

Keywords:
Novelty detection
Feature extraction
Correntropy
Half-quadratic optimization

ABSTRACT

In this paper, a novel feature extraction method based on regularized correntropy criterion (FEND-RCC) is proposed for novelty detection. In FEND-RCC, the presented criterion aims to maximize the difference between the correntropy of the normal data with their mean and the correntropy of the novel data with the mean of the normal data. Moreover, the optimal projection vectors in the objective function of FEND-RCC are iteratively obtained by the half-quadratic optimization technique. Experimental results on two synthetic data sets and thirteen benchmark data sets for novelty detection demonstrate that FEND-RCC is superior to its related approaches.

1. Introduction

Novelty detection is usually regarded as one-class classification problem [1]. The difference between one-class classification and two-class classification lies in such a way that the samples in the former problem are seriously imbalanced. There are many examples of novelty detection problems in our real world, such as fault detection [2], network intrusion detection [3], medical diagnosis [4], image segmentation [5], handwritten digit recognition [6], among others [7,8].

There are many existed novelty detection methods. Markou and Singh [9,10] provided us with a detailed survey. The commonly used novelty detection methods can be classified into three main categories below.

- *Statistical methods* [11,9]: Such as proximity-based techniques, parametric methods, non-parametric methods, and semi-parametric methods.
- *Neural network based methods* [9]: Such as multi-layer perceptron, radial basis function neural network, adaptive resonance theory models, auto-associator approaches, and Hopfield network.
- *Kernel based methods* [12]: Such as support vector machines, kernel principal component analysis, kernel Fisher discriminant analysis, and single-class minimax probability machine [13].

Besides the above methods, some new novelty detection approaches have been proposed recently. Wu and Ye [14] proposed a small sphere and large margin (SSLM) method for novelty detection. They demonstrated that SSLM outperforms one-class support vector machine [15] and support vector data description [1] on benchmark data sets. Sotiris et al. [16] proposed a novelty detection method based on one-class support vector machine and utilized Bayesian framework to estimate the posterior class probabilities of testing samples. Barakat et al. [17] proposed a self-adaptive growing neural network classifier and successfully applied it to tackle the faults detection and diagnosis problems. Angiulli [18] proposed a prototype-based domain description method for novelty detection. In comparison with the nearest neighbor-based one-class classifier, prototype-based domain description is more suitable for tackling large data sets.

However, the aforementioned novelty detection approaches suffer from a limitation that they have to utilize all the features for their training and testing. For certain novelty detection tasks, there may exist irrelevant features in the given training data. Moreover, the foresaid novelty detection methods have to face the curse of dimensionality [19]. Although they are successful for low-dimensional data sets, their accuracy and efficiency may deteriorate significantly in high-dimensional feature space. As is well known, dimensionality reduction is usually utilized to tackle the foresaid problems. The existing dimensionality reduction methods can be classified into two main categories: feature selection and feature extraction. Feature selection directly selects a smaller feature subset from the original features, while feature extraction constructs a new lower-dimensional space utilizing the information from all original

* Corresponding author.
E-mail address: hjxing@hbu.edu.cn (H.-J. Xing).

0925-2312/$ - see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2013.12.007

features. The traditional feature extraction methods were extensively studied by Ding et al. [20]. Among them, principal component analysis (PCA), linear discriminant analysis (LDA) and locality preserving projections (LPP) are three commonly used approaches.

Although PCA, LDA, and LPP have been widely applied in pattern recognition, they are very sensitive to noise. To improve the robustness of these feature extraction methods against noise, their corresponding improved versions based on information theoretic learning (ITL) [21] have been proposed. Based on the non-parametric quadratic entropy, He et al. [22,23] proposed the maximum entropy based PCA (MaxEnt-PCA) and the maximum entropy based robust discriminant analysis (MaxEnt-RDA), which are respectively regarded as the PCA and LDA from the viewpoint of ITL. Yuan and Hu [24] proposed a robust feature extraction framework based on ITL and verified that LPP is a special case within the proposed framework. Recently, the definition and properties of correntropy was introduced into ITL by Santamaria et al. [25]. As stated in [26], mean square error (MSE) is a global similarity measure whereas correntropy is a local one. Moreover, MSE is regarded as a L2-norm distance, while correntropy is considered to be L2-norm distance if samples are close, L1-norm distance if the samples become further apart, and L0-norm distance if the samples are far apart. Due to its flexibility, correntropy has been successfully utilized to design different cost functions. He et al. [27] proposed a new rotational-invariant PCA based on the maximum correntropy criterion and demonstrated that their proposed method can outperform robust rotation-invariant PCAs based on L1-norm.

To overcome the curse of dimensionality problem faced by novelty detection, feature extraction method can also be considered. Tax and Müller [28] applied PCA to novelty detection and found that low-variance directions contain more discriminant information which is different from the common usage of PCA. Villalba and Cunningham [29] empirically compared the performances of several existing dimensionality reduction methods on novelty detection. They found that LPP achieves the best performance in comparison with PCA, Laplacian score and the other related approaches when none of the input features are irrelevant. McBain and Timusk [30] proposed a feature extraction method for novelty detection (FEND) and applied it to deal with the fault detection problem. The aim of FEND is to find a subspace that maximizing the difference between the sum of squared deviations of the normal data with their mean and the sum of squared deviations of the novel data with the mean of the normal data. Moreover, FEND maximizes the variance of the projected normal data, which makes FEND regarded as the other PCA technique [30].

Inspired by correntropy and FEND, a robust feature extraction method based on the regularized correntropy criterion, namely, FEND-RCC is proposed for novelty detection. There are three main differences between the proposed method and FEND, which are listed below.

- The objective function of FEND is maximizing the difference between the sum of squared deviations of the normal data with their mean and the sum of squared deviations of the novel data with the mean of the normal data, while that of FEND-RCC is maximizing the difference between the correntropy of the normal data with their mean and the correntropy of the novel data with the mean of the normal data.
- A regularization term is added into the objective function of FEND-RCC, which makes it more robust against noise and possess higher generalization ability.
- From another point of view, FEND-RCC maximizes the within-class scatter of the projected normal data while maximizes the term like between-class scatter of the projected normal data and novel data, which makes FEND-RCC regarded as the other LDA technique.

The rest of the paper is organized as follows. The feature extraction method for novelty detection (FEND) is briefly reviewed in Section 2. In Section 3, the proposed feature extraction method based on the regularized correntropy criterion (FEND-RCC) is expatiated. Experiments to validate the proposed method are conducted in Section 4. Finally, Section 5 concludes the paper.

2. Feature extraction for novelty detection

The aim of novelty detection is to classify a testing sample as normal or novel. Most of the traditional novelty detection approaches assume that the normal data and novel data can be easily separated. However, in most cases, normal data may be effectively clustered while novel data cannot be efficiently clustered. Moreover, it is possible that the means of normal data and novel data could be completely overlapped which makes the traditional feature extraction approach become inefficient for tackling novelty detection problems. To overcome the foresaid problem, McBain and Timusk [30] proposed a feature extraction method for novelty detection named FEND.

Given N_T normal data $\{\mathbf{t}_i\}_{i=1}^{N_T}$ and N_O novel data $\{\mathbf{o}_j\}_{j=1}^{N_O}$, the mean of normal data is calculated by

$$\mathbf{m}_T = \frac{1}{N_T}\sum_{i=1}^{N_T}\mathbf{t}_i. \tag{1}$$

The optimal projection vector $\mathbf{w}^*$ of FEND can be obtained by solving the following optimization problem [30]

$$\begin{aligned} J(\mathbf{w}) &= \max_{\mathbf{w}}\left[\sum_{i=1}^{N_T}(\mathbf{w}^t\mathbf{t}_i-\mathbf{w}^t\mathbf{m}_T)^2-\sum_{j=1}^{N_O}(\mathbf{w}^t\mathbf{o}_j-\mathbf{w}^t\mathbf{m}_T)^2\right] \\ &= \max_{\mathbf{w}}\left\{\sum_{i=1}^{N_T}[\mathbf{w}^t(\mathbf{t}_i-\mathbf{m}_T)]^2-\sum_{j=1}^{N_O}[\mathbf{w}^t(\mathbf{o}_j-\mathbf{m}_T)]^2\right\} \\ &= \max_{\mathbf{w}}\left\{\mathbf{w}^t\left[\sum_{i=1}^{N_T}(\mathbf{t}_i-\mathbf{m}_T)(\mathbf{t}_i-\mathbf{m}_T)^t\right]\mathbf{w}\right. \\ &\quad \left.-\mathbf{w}^t\left[\sum_{j=1}^{N_O}(\mathbf{o}_j-\mathbf{m}_T)(\mathbf{o}_j-\mathbf{m}_T)^t\right]\mathbf{w}\right\} = \max_{\mathbf{w}}\mathbf{w}^t(\mathbf{S}_T-\mathbf{O}_T)\mathbf{w}, \end{aligned} \tag{2}$$

where $\mathbf{S}_T$ is the scatter matrix of normal data, and $\mathbf{O}_T$ is the scatter matrix of novel data.

Taking the constraint $\|\mathbf{w}\|^2=1$ into consideration, one can get the following optimization problem to obtain the optimal projection vector $\mathbf{w}^*$

$$\begin{aligned} &\max_{\mathbf{w}} \quad \mathbf{w}^T(\mathbf{S}_T-\mathbf{O}_T)\mathbf{w} \\ &\text{s.t.} \quad \|\mathbf{w}\|^2=1, \end{aligned} \tag{3}$$

where $\|\cdot\|$ denotes the L2-norm of a vector. According to the Lagrange multiplier method, the following equation can thus be obtained

$$(\mathbf{S}_T-\mathbf{O}_T)\mathbf{w}=\lambda\mathbf{w}, \tag{4}$$

where λ is the Lagrange multiplier. The k optimal projection vectors $\mathbf{w}_1^*, \mathbf{w}_2^*, \ldots, \mathbf{w}_k^*$ can be obtained by the maximum eigenvalue solutions to the generalized eigenvalue problem (4).

3. Feature extraction for novelty detection based on regularized correntropy criterion

In this section, the proposed feature extraction method for novelty detection based on regularized correntropy criterion is introduced. Moreover, its optimization method is also presented.