Neurocomputing 145 (2014) 263–268

Contents lists available at ScienceDirect

Neurocomputing

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

Fault detection based on a robust one class support vector machine

Shen Yin [a,b], Xiangping Zhu [a,*], Chen Jing [a]

[a] College of Engineering, Bohai University, Liaoning 121013, China
[b] Harbin Institute of Technology, Harbin 150001, China

ARTICLE INFO

Article history:
Received 19 February 2014
Received in revised form
29 April 2014
Accepted 12 May 2014
Communicated by Bo Shen
Available online 19 June 2014

Keywords:
Support vector machines
Outliers
One class support vector machines
Fault detection

ABSTRACT

A new fault detection scheme based on the proposed robust one class support vector machine (1-class SVM) is constructed in this paper. 1-class SVM is a special variant of the general support vector machine (SVM) and since only the normal data is required for training, 1-class SVM is widely used in anomaly detection. However, experiments show that 1-class SVM is sensitive to the outliers included in the training data set. To cope with this problem, a robust 1-class SVM is proposed in this paper. With the designed penalty factors, the robust 1-class SVM can depress the influences of outliers. Fault detection scheme is constructed based on the robust 1-class SVM. The simulation example shows that the robust 1-class SVM is superior to the general 1-class SVM, especially when the training data set is corrupted by outliers, and the fault detection scheme based on robust 1-class SVM presents satisfactory performances.

1. Introduction

Fault detection in the process industry can be defined as determining whether an abnormal process behavior has occurred. Since the need to increase reliability and to decrease the possible loss of production due to the breakdown in the underlying systems, detecting any kind of potential or occurred abnormalities in the process is extremely important. The traditional routes for fault detection are model based. Under the constructed process model, a fault detection system can be successfully designed by a large number of standard methods [1–4]. The prerequisite of these model based techniques is that the process should be well understood. However, this is difficult for the complex control system, especially for the modern large-scale process. An alternative group of methods, i.e. data-driven techniques, have emerged and received considerable attentions [5–9]. Based on large amount of data from available sensor measurements, event-logs as well as records, data-driven techniques directly construct the fault detection schemes without any prior knowledge about the underlying systems. SVM is a relatively new data mining and machine learning technique. Due to its excellent generalization ability compared with the traditional intelligent methods, e.g. neural network, SVM has been famous and popular in many areas. And a number of researches have reported the successes in applying SVM for process monitoring and fault diagnosis. In [10], SVM was used along with continuous wavelet transform to perform bearing fault detection and diagnosis of induction motor, and higher fault detection rates are obtained compared with artificial neural network. Gryllias and Antoniadis proposed a hybrid two stage one-against-all SVM approach for rolling element bearing fault detection. The simulation example gives the result that the fault classification accuracy of the new approach is satisfactory [11]. Chen et al. constructed a SVM based model to analyze the failures of the turbines in thermal power facilities and their experimental results show that SVM is superior to linear discriminant analysis and back-propagation neural networks in faults classification performance [12].

Given two class of data sets, the basic idea of SVM is establishing a hyperplane to separate these two class data sets with a maximum margin. The margin of the hyperplane can be defined as the minimum distance of the data points to hyperplane. In the case of the data sets which are not linearly separable, the 'kernel trick' is used. The data sets become linearly separable by mapping them into a higher dimensional feature space and the hyperplane is constructed in this space. In order to apply SVM for fault detection, both the normal and all kinds of faulty data sets of the industry process should be collected for training SVM. The normal operation data set is easily available, while the faulty data sets are difficult to be obtained for the reasons: (1) it is expensive to make the industry process work in the faulty condition; (2) since the fault categories are many and varied in the practical industry process, it is impossible to incorporate all of fault categories for training SVM. Moreover, whether using the one-against-one or one-against-all strategy, several or even more SVM classifiers should be trained in only one fault detection scheme and this makes the fault detection complex. 1-class SVM is a special variant of SVM and only the normal operation data set is needed for

* Corresponding author.
E-mail addresses: shen.yin2011@googlemail.com (S. Yin), xiangping.zhu2010@gmail.com (X. Zhu), jingchen8855@gmail.com (C. Jing).

http://dx.doi.org/10.1016/j.neucom.2014.05.035
0925-2312/

training. It devotes to constructing a decision boundary which has the maximum margin between the normal data points and the origin [13]. If a new sample locates within the boundary, it will be classified as a normal operation point. In contrast, it will be labeled as abnormality when it lies outside of the boundary. Since no faulty data is needed for training, 1-class SVM is easier to be trained and more applicable for fault detection than standard SVM.

Outliers are the instances which deviate from the majority of the data points. In practical data set, the outliers always exist for the reasons such as instrument failure, formatting errors and non-representative sampling. Unfortunately, experiments show that 1-class SVM is sensitive to outliers included in the training data set. To cope with this problem, a robust 1-class SVM is proposed in this paper. Based on the Euclidean distance of every normal data point to the center of the data set, the adaptive penalty factors are designed and they make outliers have less influences on the decision boundary of 1-class SVM. In order to implement 1-class SVM for fault detection, a distance metric and the corresponding threshold are also introduced. Finally, a simulation example is utilized to demonstrate the effectiveness of the robust 1-class SVM.

The reminder of this paper is structured as follows. In Section 2, the general 1-class SVM is reviewed and the problem that the general 1-class SVM is sensitive to outliers is formulated. Section 3 introduces the robust 1-class SVM. Section 4 analyzes how to implement the robust 1-class SVM for fault detection and the simulation example is presented in Section 5. Finally, the conclusions are made in the last section.

2. Preliminaries and problem formulation

2.1. Algorithm of the general 1-class SVM

Different from the general SVM, 1-class SVM that proposed by Schölkopf et al. is to build a decision boundary which has the maximum margin between the normal data set and the origin [13]. Considering a data set $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_N]^T\in\mathcal{R}^{N\times M}$ which denotes the normal data set, in order to obtain the boundary, a optimization model is considered as follows:

$$\begin{aligned}&\min_{\mathbf{w}\in F,\boldsymbol{\xi}\in\mathcal{R}^N,\rho\in\mathcal{R}} \quad \tfrac{1}{2}\|\mathbf{w}\|^2+\tfrac{1}{N\nu}\sum_{i=1}^{N}\xi_i-\rho\\ &\text{subject to} \quad \mathbf{w}\cdot\boldsymbol{\Phi}(\mathbf{x}_i)\geq\rho-\xi_i,\quad \xi_i\geq 0.\end{aligned} \tag{1}$$

where N is the number of the data points, ν is a regularization parameter and ξ_i is the slack variable for point $\mathbf{x}_i$ that allow it to locate outside of the decision boundary and $\boldsymbol{\xi}=[\xi_1,\ldots,\xi_N]$. $\mathbf{w}$ and ρ are the parameters which determine the decision boundary and they are target variables of the optimization problem. The decision boundary can be formulated as

$$f(\mathbf{x})=\mathbf{w}\cdot\boldsymbol{\Phi}(\mathbf{x})-\rho \tag{2}$$

where $\mathbf{x}\in\mathcal{R}^M$, and $\boldsymbol{\Phi}$ is a kind of mapping. For the classification problem of two categories, the data sets are not always linearly separable in the original space. $\boldsymbol{\Phi}$ projects the original data sets into a higher dimensional space which is the so-called feature space and the non-separable data sets become linearly separable in this space. However, $\boldsymbol{\Phi}$ is inexplicit in the practical application and only the dot product form $\boldsymbol{\Phi}(\mathbf{x}_i)\cdot\boldsymbol{\Phi}(\mathbf{x}_j)$ is necessary to be known. $\boldsymbol{\Phi}(\mathbf{x}_i)\cdot\boldsymbol{\Phi}(\mathbf{x}_j)$ is the kernel function and K is usually used to represent it, i.e. $K(\mathbf{x}_i,\mathbf{x}_j)=\boldsymbol{\Phi}(\mathbf{x}_i)\cdot\boldsymbol{\Phi}(\mathbf{x}_j)$. The commonly used kernel functions are listed as

- Linear kernel: $K(\mathbf{x}_i,\mathbf{x}_j)=\mathbf{x}_i^T\mathbf{x}_j$
- Polynomial kernel: $K(\mathbf{x}_i,\mathbf{x}_j)=(\gamma\mathbf{x}_i^T\mathbf{x}_j+c)^p$
- Radial basis function kernel: $K(\mathbf{x}_i,\mathbf{x}_j)=e^{-\|\mathbf{x}_i-\mathbf{x}_j\|^2/2\sigma^2}$
- Sigmoidal kernel: $K(\mathbf{x}_i,\mathbf{x}_j)=\tanh(k\mathbf{x}_i^T\mathbf{x}_j-c)$

where γ and c are constants, σ is the width of radial basis function (rbf) kernel and p is the degree of the polynomial. Keerthi et al. pointed out that the rbf kernel can approximate most kernel functions if the parameter σ is chosen appropriately [14]. Moreover, there exists only one parameter in rbf kernel which can make the kernel easily tuned. Thus, rbf kernel is selected as the kernel function $K(\mathbf{x}_i,\mathbf{x}_j)$ in this paper.

In order to solve the optimization problem (1), Lagrange multipliers $\alpha_i\geq 0$ and $\beta_i\geq 0$ $(i=1,\ldots,N)$ are introduced and the Lagrange equation is formed as

$$\begin{aligned}L(\mathbf{w},\boldsymbol{\xi},\rho,\boldsymbol{\alpha},\boldsymbol{\beta})=&\frac{1}{2}\|\mathbf{w}\|+\frac{1}{N\nu}\sum_{i=1}^{N}\xi_i-\rho\\ &-\sum_{i=1}^{N}\alpha_i(\mathbf{w}\cdot\boldsymbol{\Phi}(x_i)-\rho+\xi_i)-\sum_{i=1}^{N}\beta_i\xi_i\end{aligned} \tag{3}$$

The partial derivatives of the Lagrangian equation with respect to $\mathbf{w}$, $\boldsymbol{\xi}$ and ρ are set to zero. Then, $\mathbf{w}$ and $\boldsymbol{\alpha}$ can be formulated as

$$\mathbf{w}=\sum_{i=1}^{N}\alpha_i\boldsymbol{\Phi}(x_i) \tag{4}$$

$$\alpha_i=\frac{1}{N\nu}-\beta_i \quad \sum_{i=1}^{N}\alpha_i=1 \tag{5}$$

Substitute (4)–(5) into Lagrangian Eq. (3) and its dual form is presented as

$$\begin{aligned}&\min_{\boldsymbol{\alpha}} \quad \boldsymbol{\alpha}^T\mathbf{H}\boldsymbol{\alpha}\\ &\text{subject to} \quad 0\leq\alpha_i\leq\frac{1}{N\nu},\quad \sum_{i=1}^{N}\alpha_i=1\end{aligned} \tag{6}$$

where $\boldsymbol{\alpha}=[\alpha_1\ldots\alpha_N]^T$. According to the Kuhn–Tucker conditions, the data points can be classified into three categories: (1) the points with $\alpha_i=0$ locate within the boundary; (2) the points with $0<\alpha_i<1/N\nu$ are on the boundary and the corresponding ξ_i equal to zero; and (3) the points which satisfy $\alpha_i=1/N\nu$ fall outside of the boundary. The points with $\alpha_i>0$ are the so-called support vectors. $\mathbf{H}$ is the kernel matrix and the factor of $\mathbf{H}$, i.e. H_{ij}, can be expressed as

$$H_{ij}=K(\mathbf{x}_i,\mathbf{x}_j)=\boldsymbol{\Phi}(\mathbf{x}_i)\cdot\boldsymbol{\Phi}(\mathbf{x}_j) \tag{7}$$

Solve the optimization problem (6) to get $\boldsymbol{\alpha}$ and then ρ can be given as

$$\rho=\frac{1}{n_s}\sum_{i=1}^{n_s}\sum_{j=1}^{N}\alpha_jK(\mathbf{x}_i,\mathbf{x}_j)\alpha \tag{8}$$

where n_s is the number of support vectors which satisfy $\xi_i=0$ and $0<\alpha_i<1/N\nu$.

Tax et al. proposed another form of 1-class SVM which is called support vector data description (SVDD) [15]. The basic idea of SVDD is to construct a minimum-volume hypersphere in a high dimensional feature space to enclose as much as normal data points. Both of these two forms of 1-class SVM which are respectively proposed by Schölkopf et al. and Tax et al. have an equivalent solution if the diagonal entries of kernel matrix $\mathbf{H}$ equal to a constant [13].

2.2. Problem formulation

In the practical application, the target data set always includes some outliers for the reasons such as instrument failure, formatting errors and non-representative sampling. For 1-class SVM, it allows some data points to locate outside of the boundary by introducing in the slack variables ξ_i $(i=1,\ldots,N)$. And the number of points located outside of the boundary can be controlled by the penalty factor $1/N\nu$. The smaller the penalty factor is, the more possibilities for data points to locate outside of the decision