Hindawi Publishing Corporation
Mathematical Problems in Engineering
Volume 2016, Article ID 1658758, 12 pages
http://dx.doi.org/10.1155/2016/1658758

Research Article

A Core Set Based Large Vector-Angular Region and Margin Approach for Novelty Detection

Jiusheng Chen, Xiaoyu Zhang, and Kai Guo

College of Electronics, Information & Automation, Civil Aviation University of China, Tianjin 300300, China

Correspondence should be addressed to Jiusheng Chen; jschen@cauc.edu.cn

Received 2 November 2015; Revised 10 January 2016; Accepted 12 January 2016

Academic Editor: Muhammad N. Akram

A large vector-angular region and margin (LARM) approach is presented for novelty detection based on imbalanced data. The key idea is to construct the largest vector-angular region in the feature space to separate normal training patterns; meanwhile, maximize the vector-angular margin between the surface of this optimal vector-angular region and abnormal training patterns. In order to improve the generalization performance of LARM, the vector-angular distribution is optimized by maximizing the vector-angular mean and minimizing the vector-angular variance, which separates the normal and abnormal examples well. However, the inherent computation of quadratic programming (QP) solver takes $O(n^3)$ training time and at least $O(n^2)$ space, which might be computational prohibitive for large scale problems. By $(1+\varepsilon)$ and $(1-\varepsilon)$-approximation algorithm, the core set based LARM algorithm is proposed for fast training LARM problem. Experimental results based on imbalanced datasets have validated the favorable efficiency of the proposed approach in novelty detection.

1. Introduction

The task of novelty detection is to learn a model from normal examples in training patterns and hence can classify the test patterns. In real-world novelty detection applications, it is usually assumed that normal training patterns can be well sampled, while abnormal training patterns are severely undersampled, which is due to expensive measurement cost or infrequency of abnormal events. Therefore, only normal training patterns are used to build detection model in most novelty detection algorithms. Generally, novelty detection may be seen as one-class classification problem. Recently, novelty detection has gained much research attention in real-world applications such as network intrusion detection [1], jet engine health monitoring [2], medical data [3], and aviation safety [4, 5].

In this paper, the kernel-based novelty detection algorithm is studied in-depth, which is very popular and has been proved to be successful recently. Various kernel-based novelty detection approaches have been proposed, such as one-class support vector machine (OCSVM) [6] and support vector data description (SVDD) [7]. OCSVM was proposed by Schölkopf et al. [6], in which, to improve generalization ability, novelty detection boundary is constructed to separate the origin from the input samples with the maximal margin. The performance of OCSVM is very sensitive to the parameters, making it difficult to be generalized to other applications [8].

SVDD was proposed by Tax and Duin [7], in which the minimal ball is constructed to enclose most of the training samples. Novelty point is assessed by determining whether a test point lies within the minimal ball or not. The margin between the closed boundary surrounding the positive data and that surrounding the negative data is zero, which makes the method of poor generalization ability. A small sphere and large margin (SSLM) approach was proposed by Wu and Ye [9], in which the smallest hypersphere is constructed to surround the normal data; meanwhile, the margin from any outlier to this hypersphere is as large as possible. An incremental weighted one-class support vector machine for mining streaming data was proposed by Krawczyk and Wózniak [10, 11], in which the weights to each object are modified according to its level of significance, and the shape of the decision boundary is influenced only by new objects that carry new and useful knowledge extending the competence of the classifier.

Support vector machine (SVM) can be solved through figuring out quadratic programming (QP) problem, which has the important computational advantage of avoiding the problem of local minima. However, solving the corresponding SVM problems using the naive implementation of QP solver takes $O(n^3)$ computational time complexity and at least $O(n^2)$ space complexity if the number of training patterns is n. Obviously, the naive implementation of QP solver is difficult to meet the practical application of novelty detection in large scale datasets. Tsang et al. proposed the core vector machine (CVM) [12, 13] as the approximation algorithm of minimum enclosing ball (MEB) for large scale problems. The key idea is that the implementation of QP solver for corresponding SVM problems could be equivalently viewed as MEB problems. By utilizing an approximation algorithm for the MEB problem in computational geometry, the time complexity of CVM algorithm is linear to the number of training patterns. Moreover, the space complexity is irrelevant to the number of training patterns.

As mentioned above, only normal training patterns are used to build the detection model in most novelty detection algorithms. In practical applications of novelty detection, it is difficult, but not impossible, to obtain a very few abnormal training patterns. For instance, in machine fault detection, in addition to extensive measurements on the normal working conditions, there may be also some measurements on faulty situations [14]. Recently, extensive and comprehensive researches have been carried out in both academia and industry to solve the imbalanced novelty detection problem.

Kernel-based novelty detection based on imbalanced data is researched in this paper. Suppose $S = \{(\mathbf{x}_i, y_i)\}, i = 1, \ldots, n$, is a given training dataset with n examples, where $\mathbf{x}_i \subset R^d$ is the ith input instance, $y_i \in \{-1, +1\}$ is a class identity label associated with instance $\mathbf{x}_i$, $S_{\text{maj}} \subset S$ is the set of majority training patterns and $|S_{\text{maj}}| = m_1$, $S_{\text{min}} \subset S$ is the set of minority training patterns and $|S_{\text{min}}| = m_2$, and $m_1 + m_2 = n$. $\phi(\cdot)$ is the feature mapping function defined by a given kernel function $\kappa(\cdot, \cdot)$. The length of the perpendicular projection of the training pattern $\phi(\mathbf{x}_i)$ onto the vector $\mathbf{v}$ is expressed as $\langle \mathbf{v}, \phi(\mathbf{x}_i) \rangle$, which actually reflects the information about the angular and the Euclidean distances between $\mathbf{v}$ and $\phi(\mathbf{x}_i)$ in the Euclidean vector space. According to the definition in [15], $\langle \mathbf{v}, \phi(\mathbf{x}_i) \rangle$ is called vector-angular.

In this paper, a large vector-angular region and margin (LARM) algorithm and its fast training method based on core set are proposed for novelty detection, where the training patterns are imbalanced. The main contributions of this paper lie in three aspects. Firstly, the boundary of SVM is only determined by the support vectors and the distribution of the data in the training set is not considered [16]. However, recent theoretical results have proved that data distribution information is crucial to the generalization performance [17, 18]. The proposed algorithm in this paper aims to find an optimal vector $\mathbf{v}$ in the feature space, in which the mean and the variance of vector-angular are maximized and minimized, respectively. Therefore, normal and abnormal examples are well separated when projected onto the optimal vector $\mathbf{v}$ joining their large mean and small variance. Secondly, the proposed LARM integrates one-class and binary classification algorithms to tackle the novelty detection problem based on imbalanced data, which constructs the largest vector-angular region in the feature space to separate normal training patterns and maximizes the vector-angular margin between the optimal vector-angular region and the abnormal data. Since the number of normal training patterns is sufficient, the largest vector-angular region is constructed accurately, which can minimize the chance of accepting the normal examples. To achieve better generalization performance, the vector-angular margin between the surface of this optimal vector-angular region and the abnormal data is maximized. Thirdly, the core set based LARM algorithm is proposed for fast training LARM problem. The time and space complexity of core set based LARM are linear to and independent of the number of training patterns, respectively.

The structure of this paper is organized as follows. Section 1 introduces the novelty detection technique and presents an analysis of the existing problems. Section 2 introduces ν-support vector machine (ν-SVM), two-class SVDD, and maximum vector-angular margin classifier (MAMC). Section 3 presents the proposed LARM for novelty detection and its fast training method based on core set. Experimental results are shown in Section 4 and conclusions are given in Section 5.

2. ν-SVM, SVDD, and MAMC

2.1. ν-SVM. ν-SVM was proposed by Schölkopf et al. [19] to solve the binary classification problem, which uses the parameter ν to control the number of support vectors and the bound of the classification errors. ν-SVM can be modeled as follows:

$$\begin{aligned} \min_{\mathbf{w},\rho,b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 - \nu\rho + \frac{1}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i\left(\mathbf{w}^{\mathrm{T}}\phi\left(\mathbf{x}_i\right) + b\right) \geq \rho - \xi_i, \quad i = 1, \ldots, n \\ & \xi_i \geq 0, \quad i = 1, \ldots, n, \\ & \rho > 0, \end{aligned} \tag{1}$$

where $\mathbf{w}$ is the normal vector of the decision hyperplane, b is the bias of the classifier, ρ is the margin, $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_n]^{\mathrm{T}}$ is the vector of slack variables, and ν is a positive constant. ν-SVM obtains the optimal hyperplane $\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) + b = 0$ for separating the two classes with a maximal margin $\rho/(2\|\mathbf{w}\|)$. To classify a testing instance $\mathbf{z} \in R^d$, the decision function takes the sign function of the optimal hyperplane $f(\mathbf{z}) = \text{sgn}(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{z}) + b)$.

2.2. SVDD. One-class SVDD and two-class SVDD were proposed by Tax and Duin in 2004 [7], in which the minimal ball is constructed to enclose most of the training patterns. Here, we only review two-class SVDD that can utilize the abnormal data. Two-class SVDD can be modeled as follows:

$$\begin{aligned}\min_{R,\mathbf{c},\boldsymbol{\xi}} \quad & R^2 + C_1 \sum_{i=1}^{m_1} \xi_i + C_2 \sum_{j=m_1+1}^{n} \xi_j \\ \text{s.t.} \quad & \left\| \phi(\mathbf{x}_i) - \mathbf{c} \right\|^2 \le R^2 + \xi_i, \quad i = 1, \dots, m_1 \\ & \left\| \phi(\mathbf{x}_j) - \mathbf{c} \right\|^2 \ge R^2 - \xi_j, \quad j = m_1 + 1, \dots, n \\ & \xi_k \ge 0, \quad k = 1, \dots, n,\end{aligned} \tag{2}$$

where R and $\mathbf{c}$ are the radius and the center of the hypersphere, C_1 and C_2 are two trade-off parameters which can treat imbalanced datasets, and $\boldsymbol{\xi} = [\xi_1, \dots, \xi_n]^{\mathrm{T}}$ is the vector of slack variables. The testing instance $\mathbf{z} \in R^d$ can be determined, whether it is inside of the optimal hypersphere or not. Hence, the decision function of two-class SVDD is $f(\mathbf{z}) = \operatorname{sgn}(R^2 - \|\phi(\mathbf{z}) - \mathbf{c}\|^2)$.

2.3. MAMC. MAMC was proposed by Hu et al. in 2012 [15], which attempts to find an optimal vector $\mathbf{v}$ in the feature space based on the maximum vector-angular margin. MAMC can be modeled as follows:

$$\begin{aligned}\min_{\rho,\mathbf{v},\boldsymbol{\xi}} \quad & -\nu\rho + \frac{1}{n}\sum_{i=1}^{n} \left\| \phi(\mathbf{x}_i) - \mathbf{v} \right\|^2 + \frac{C}{n}\sum_{i=1}^{n}\xi_i \\ \text{s.t.} \quad & y_i \phi(\mathbf{x}_i)^{\mathrm{T}} \mathbf{v} \ge \rho - \xi_i, \quad i = 1, \dots, n \\ & \xi_i \ge 0, \quad i = 1, \dots, n,\end{aligned} \tag{3}$$

where $\mathbf{v}$ is the optimized vector, ρ is the vector-angular margin, $\boldsymbol{\xi} = [\xi_1, \dots, \xi_n]^{\mathrm{T}}$ is the vector of slack variables, and C and ν are two positive constants. To classify a testing instance $\mathbf{z} \in R^d$, the decision function is defined as $f(\mathbf{z}) = \operatorname{sgn}(\sum_{i=1}^{n}(1/n + \alpha_i y_i/2)\kappa(\mathbf{x}_i, \mathbf{z}))$.

3. Core Set Based Large Vector-Angular Region and Margin

In this section, LARM algorithm and its fast training method based on core set are proposed for novelty detection with imbalanced data.

3.1. LARM. To tackle the novelty detection problem on imbalanced data, the distribution of vector-angular and maximization of vector-angular margin are considered in this paper. Figure 1 illustrates the principle of LARM.

Firstly, LARM is adopted to find an optimal vector $\mathbf{v}$ in the feature space, which attempts to maximize the vector-angular mean and minimize the vector-angular variance simultaneously. Here, the vector-angular expresses the length of projection of training pattern $\phi(\mathbf{x}_i)$ onto the optimal vector $\mathbf{v}$. Therefore, the normal and abnormal examples are well separated when projected onto the optimal vector $\mathbf{v}$ joining their large mean and small variance.

Secondly, for the learning problem on imbalanced data, the largest vector-angular region in the feature space is constructed to separate the normal data. Since the number of normal training patterns is sufficient, the largest vector-angular region is constructed accurately, which can minimize the chances of accepting the normal examples. Meanwhile, to achieve a favorable generalization performance, the vector-angular margin between the surface of this optimal vector-angular region and the abnormal data is maximized.

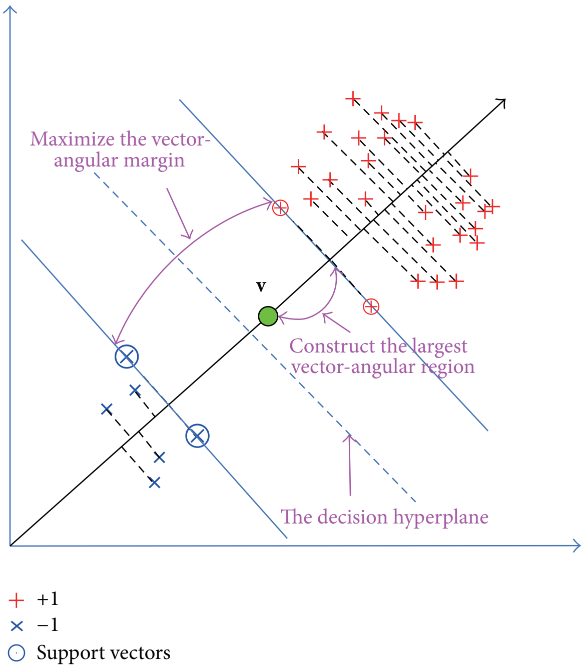

Maximize the vector-angular margin
v
Construct the largest vector-angular region
The decision hyperplane
+ +1
× −1
○ Support vectors

Figure 1: The schematic illustration of LARM principle.

3.1.1. Primal Formulation of LARM. Formally, define the training pattern matrix $\mathbf{X} = [\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)]$, label column vector $\mathbf{y} = [y_1, \dots, y_n]^{\mathrm{T}}$, and label diagonal matrix $\mathbf{Y} = \operatorname{diag}(y_1, \dots, y_n)$. According to the definition in [18], the vector-angular mean $\bar{\gamma}$ and vector-angular variance $\hat{\gamma}$ between training patterns $(\phi(\mathbf{x}_i), y_i)$, $i = 1, \dots, n$ and vector $\mathbf{v}$ can be expressed as

$$\begin{aligned}\bar{\gamma} &= \frac{1}{n}\sum_{i=1}^{n} y_i \mathbf{v}^{\mathrm{T}} \phi(\mathbf{x}_i) = \frac{1}{n}(\mathbf{X}\mathbf{y})^{\mathrm{T}}\mathbf{v}, \\ \hat{\gamma} &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(y_i \mathbf{v}^{\mathrm{T}}\phi(\mathbf{x}_i) - y_j \mathbf{v}^{\mathrm{T}}\phi(\mathbf{x}_j)\right)^2 \\ &= \frac{2}{n^2}\left(n\mathbf{v}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{v} - \mathbf{v}^{\mathrm{T}}\mathbf{X}\mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{v}\right).\end{aligned} \tag{4}$$

Then, the primal LARM can be formulated as the following optimization problem:

$$\begin{aligned}\min_{\omega,\rho,\mathbf{v},\boldsymbol{\xi}} \quad & -\omega^2 - \nu\rho^2 + \frac{2\lambda}{n^2}\left(n\mathbf{v}^{\mathrm{T}}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{v} - \mathbf{v}^{\mathrm{T}}\mathbf{X}\mathbf{y}\mathbf{y}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{v}\right) \\ & - \frac{\lambda}{n}(\mathbf{X}\mathbf{y})^{\mathrm{T}}\mathbf{v} + \frac{1}{\nu_1 m_1}\sum_{i=1}^{m_1}\xi_i + \frac{1}{\nu_2 m_2}\sum_{j=m_1+1}^{n}\xi_j\end{aligned}$$

$$\begin{aligned}\text{s.t.}\quad & \mathbf{v}^{\mathrm{T}}\phi\left(\mathbf{x}_i\right) \geq \omega^2 - \xi_i, \quad i = 1,\ldots,m_1 \\ & \mathbf{v}^{\mathrm{T}}\phi\left(\mathbf{x}_j\right) \leq \omega^2 - \rho^2 + \xi_j, \quad j = m_1+1,\ldots,n \\ & \xi_i \geq 0, \quad i = 1,\ldots,n,\end{aligned} \tag{5}$$

where $\mathbf{v}$ is the optimal vector, ω^2 ($\omega > 0$) is the width of vector-angular region, ρ^2 ($\rho > 0$) is the vector-angular margin, $\boldsymbol{\xi} = [\xi_1,\ldots,\xi_n]^{\mathrm{T}}$ is the vector of slack variables, and ν, ν_1, ν_2, and λ are four positive constants.

According to [18], $\mathbf{v}^*$ for problem (5) is expressed as follows:

$$\mathbf{v}^* = \sum_{i=1}^{n}\alpha_i\phi\left(\mathbf{x}_i\right) = \mathbf{X}\boldsymbol{\alpha}. \tag{6}$$

Hence, $\mathbf{X}^{\mathrm{T}}\mathbf{v} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha} = \mathbf{K}\boldsymbol{\alpha}$ can be obtained, where $\mathbf{K} = \mathbf{X}^{\mathrm{T}}\mathbf{X}$ is the kernel matrix. Problem (5) can be formulated as follows:

$$\begin{aligned}\min_{\omega,\rho,\boldsymbol{\alpha},\boldsymbol{\xi}}\quad & -\omega^2 - \nu\rho^2 + \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Q}\boldsymbol{\alpha} + \mathbf{q}^{\mathrm{T}}\boldsymbol{\alpha} + \frac{1}{\nu_1 m_1}\sum_{i=1}^{m_1}\xi_i \\ & + \frac{1}{\nu_2 m_2}\sum_{j=m_1+1}^{n}\xi_j \\ \text{s.t.}\quad & \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}_{:i} \geq \omega^2 - \xi_i, \quad i = 1,\ldots,m_1 \\ & \boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}_{:j} \leq \omega^2 - \rho^2 + \xi_j, \quad j = m_1+1,\ldots,n \\ & \xi_i \geq 0, \quad i = 1,\ldots,n,\end{aligned} \tag{7}$$

where $\mathbf{Q} = 4\lambda(n\mathbf{K}^{\mathrm{T}}\mathbf{K} - (\mathbf{K}\mathbf{y})(\mathbf{K}\mathbf{y})^{\mathrm{T}})/n^2$, $\mathbf{q} = -\lambda(\mathbf{K}\mathbf{y})/n$, and $\mathbf{K}_{:i}$ is the ith column of $\mathbf{K}$.

3.1.2. Dual Problem. To investigate the problem with constraints described as (7), the Lagrangian function is constructed as follows:

$$\begin{aligned}& L\left(\omega,\rho,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\eta}\right) \\ & = -\omega^2 - \nu\rho^2 + \frac{1}{2}\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{Q}\boldsymbol{\alpha} + \mathbf{q}^{\mathrm{T}}\boldsymbol{\alpha} + \frac{1}{\nu_1 m_1}\sum_{i=1}^{m_1}\xi_i \\ & + \frac{1}{\nu_2 m_2}\sum_{j=m_1+1}^{n}\xi_j - \sum_{i=1}^{m_1}\beta_i\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}_{:i} - \omega^2 + \xi_i\right) \\ & + \sum_{j=m_1+1}^{n}\beta_j\left(\boldsymbol{\alpha}^{\mathrm{T}}\mathbf{K}_{:j} - \omega^2 + \rho^2 - \xi_j\right) - \sum_{i=1}^{n}\eta_i\xi_i,\end{aligned} \tag{8}$$

where $\boldsymbol{\beta} = [\beta_1,\ldots,\beta_n]^{\mathrm{T}}$ and $\boldsymbol{\eta} = [\eta_1,\ldots,\eta_n]^{\mathrm{T}}$ are Lagrange multipliers. The following equations can be obtained by making the partial derivatives of $L(\omega,\rho,\boldsymbol{\alpha},\boldsymbol{\xi},\boldsymbol{\beta},\boldsymbol{\eta})$ with respect to the primal variables to zero:

$$\begin{aligned}\frac{\partial L}{\partial\omega} = 0 & \Longrightarrow -2\omega + 2\omega\sum_{i=1}^{m_1}\beta_i - 2\omega\sum_{j=m_1+1}^{n}\beta_j = 0 \\ & \Longrightarrow 2\omega\left(-1 + \sum_{i=1}^{n}\beta_i y_i\right) = 0\end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial L}{\partial\rho} = 0 & \Longrightarrow -2\nu\rho + 2\rho\sum_{j=m_1+1}^{n}\beta_j = 0 \\ & \Longrightarrow 2\rho\left(\sum_{j=m_1+1}^{n}\beta_j - \nu\right) = 0\end{aligned} \tag{10}$$

$$\frac{\partial L}{\partial\xi_i} = 0 \Longrightarrow \frac{1}{\nu_1 m_1} - \beta_i - \eta_i = 0, \quad i = 1,\ldots,m_1 \tag{11}$$

$$\frac{\partial L}{\partial\xi_j} = 0 \Longrightarrow \frac{1}{\nu_2 m_2} - \beta_j - \eta_j = 0, \quad j = m_1+1,\ldots,n \tag{12}$$

$$\begin{aligned}\frac{\partial L}{\partial\boldsymbol{\alpha}} = 0 & \Longrightarrow \mathbf{Q}\boldsymbol{\alpha} + \mathbf{q} - \sum_{i=1}^{m_1}\beta_i\mathbf{K}_{:i} + \sum_{j=m_1+1}^{n}\beta_j\mathbf{K}_{:j} = 0 \\ & \Longrightarrow \mathbf{Q}\boldsymbol{\alpha} + \mathbf{q} - \sum_{i=1}^{n}\beta_i y_i\mathbf{K}_{:i} = 0.\end{aligned} \tag{13}$$

Substituting (9)–(13) into (8), the dual form can be obtained, which omits constants without influence on optimization:

$$\begin{aligned}\max_{\boldsymbol{\beta}}\quad & -\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\mathbf{H}\boldsymbol{\beta} + \mathbf{p}^{\mathrm{T}}\boldsymbol{\beta} \\ \text{s.t.}\quad & 0 \leq \beta_i \leq \frac{1}{\nu_1 m_1}, \quad i = 1,\ldots,m_1 \\ & 0 \leq \beta_i \leq \frac{1}{\nu_2 m_2}, \quad i = m_1+1,\ldots,n \\ & \sum_{i=1}^{n}\beta_i y_i = 1 \\ & \sum_{i=1}^{n}\beta_i = 1 + 2\nu,\end{aligned} \tag{14}$$

where $\mathbf{H} = \mathbf{Y}\mathbf{K}\mathbf{Q}^{-1}\mathbf{K}\mathbf{Y}$, $\mathbf{p} = -\lambda(\mathbf{H}\mathbf{e})/n$, and $\mathbf{Q}^{-1}$ is the inverse matrix of $\mathbf{Q}$ and $\mathbf{e} = [1,\ldots,1]^{\mathrm{T}}$.

The dual problem (14) is a QP problem, which has the same form as the dual of the ν-SVM [19, 20]. Therefore, the QP problem (14) can be easily solved by SMO algorithm in LIBSVM [21].

Suppose $\boldsymbol{\beta}^*$ is the optimal vector of the dual problem (14). According to (13), $\boldsymbol{\alpha}^*$ can be expressed as follows:

$$\boldsymbol{\alpha}^* = \mathbf{Q}^{-1}\left(\sum_{i=1}^{n}\beta_i^* y_i\mathbf{K}_{:i} - \mathbf{q}\right) = \mathbf{Q}^{-1}\left(\mathbf{K}\mathbf{Y}\boldsymbol{\beta}^* - \mathbf{q}\right). \tag{15}$$

To compute ω^2 and ρ^2, two sets are considered:

$$S_1 = \left\{0 < \beta_i < \frac{1}{\nu_1 m_1},\ i = 1, \ldots, m_1\right\},$$
$$S_2 = \left\{0 < \beta_j < \frac{1}{\nu_2 m_2},\ j = m_1 + 1, \ldots, n\right\}. \tag{16}$$

According to the Karush-Kuhn-Tucker (KKT) conditions

$$\beta_i \left(\mathbf{v}^{\mathrm{T}} \phi\left(\mathbf{x}_i\right) - \omega^2 + \xi_i\right) = 0, \quad i = 1, \ldots, m_1,$$
$$\beta_j \left(\mathbf{v}^{\mathrm{T}} \phi\left(\mathbf{x}_j\right) - \omega^2 + \rho^2 - \xi_i\right) = 0, \quad j = m_1 + 1, \ldots, n, \tag{17}$$
$$\eta_i \xi_i = 0, \quad i = 1, \ldots, n,$$

and (11) and (12), $\eta_i > 0$, $\xi_i = 0$, $\eta_j > 0$, and $\xi_j = 0$ can be obtained. Hence, set $n_1 = |S_1|$ and $n_2 = |S_2|$, and ω^2 and ρ^2 can be expressed as

$$\omega^2 = \frac{1}{n_1} \sum_{\mathbf{x}_i \in S_1} \mathbf{v}^{\mathrm{T}} \phi\left(\mathbf{x}_i\right) = \frac{1}{n_1} \sum_{\mathbf{x}_i \in S_1} \sum_{k=1}^{n} \alpha_k^* \phi\left(\mathbf{x}_k\right)^{\mathrm{T}} \phi\left(\mathbf{x}_i\right)$$
$$= \frac{1}{n_1} \sum_{\mathbf{x}_i \in S_1} \sum_{k=1}^{n} \alpha_k^* \kappa\left(\mathbf{x}_i, \mathbf{x}_k\right),$$
$$\rho^2 = \frac{1}{n_1} \sum_{\mathbf{x}_i \in S_1} \mathbf{v}^{\mathrm{T}} \phi\left(\mathbf{x}_i\right) - \frac{1}{n_2} \sum_{\mathbf{x}_j \in S_2} \mathbf{v}^{\mathrm{T}} \phi\left(\mathbf{x}_j\right)$$
$$= \frac{1}{n_1} \sum_{\mathbf{x}_i \in S_1} \sum_{k=1}^{n} \alpha_k^* \phi\left(\mathbf{x}_k\right)^{\mathrm{T}} \phi\left(\mathbf{x}_i\right) \tag{18}$$
$$- \frac{1}{n_2} \sum_{\mathbf{x}_j \in S_2} \sum_{k=1}^{n} \alpha_k^* \phi\left(\mathbf{x}_k\right)^{\mathrm{T}} \phi\left(\mathbf{x}_j\right)$$
$$= \frac{1}{n_1} \sum_{\mathbf{x}_i \in S_1} \sum_{k=1}^{n} \alpha_k^* \kappa\left(\mathbf{x}_i, \mathbf{x}_k\right) - \frac{1}{n_2} \sum_{\mathbf{x}_j \in S_2} \sum_{k=1}^{n} \alpha_k^* \kappa\left(\mathbf{x}_j, \mathbf{x}_k\right).$$

3.1.3. Decision Function. It can be seen that minimizing the cost function (5) will make the width of vector-angular region ω^2 and vector-angular margin ρ^2 as large as possible. Meanwhile, the optimal vector $\mathbf{v}$ in feature space is found, which makes the normal and abnormal examples well separated when projected onto the optimal vector $\mathbf{v}$ joining their large mean and small variance. Therefore, the testing patterns can be classified in terms of the vector-angular between the vector $\mathbf{v}$ and the training patterns $\phi(\mathbf{x})$. The optimal separating hyperplane of SVM is $\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b = 0$, which is at the middle of the margin. Similarly, the separating hyperplane of LARM is defined at the center of the margin. Hence, for testing instance $\mathbf{z} \in R^d$, the decision function is expressed as follows:

$$f(\mathbf{z}) = \operatorname{sgn}\left(\mathbf{v}^{\mathrm{T}} \phi(\mathbf{z}) - \omega^2 + \frac{\rho^2}{2}\right)$$
$$= \operatorname{sgn}\left(\sum_{i=1}^{n} \alpha_i^* \kappa\left(\mathbf{x}_i, \mathbf{z}\right) - \omega^2 + \frac{\rho^2}{2}\right). \tag{19}$$

3.1.4. ν-Property. Let m^+ and m^- represent the number of margin errors of the normal and abnormal training patterns and s^+ and s^- denote the number of support vectors of the normal and abnormal training patterns, respectively. According to (9) and (10), the following formulas can be obtained:

$$\sum_{i=1}^{m_1} \beta_i = \nu + 1,$$
$$\sum_{j=m_1+1}^{n} \beta_j = \nu. \tag{20}$$

By using similar proof about ν-property in [19] and by making use of (20), inequalities (21) can be obtained:

$$\frac{m^+}{m_1} \le \nu_1 (\nu + 1) \le \frac{s^+}{m_1},$$
$$\frac{m^-}{m_2} \le \nu_2 \nu \le \frac{s^-}{m_2}. \tag{21}$$

The inequalities (21) indicate that $\nu_1(\nu + 1)$ (or $\nu_2 \nu$) is a lower bound of the fraction of support vectors in the normal (or abnormal) dataset and an upper bound of the fraction of misclassified patterns in the normal (or abnormal) dataset. The ν-property of LARM can be used for parameter selection in the following experiments.

3.2. Core Set Based LARM. As mentioned above, the dual problem of LARM can be actually formulated as a QP problem. So, solving the corresponding QP problem of LARM takes $O(n^3)$ computational time complexity and $O(n^2)$ space complexity. When the number of training patterns is large, it is thus computationally infeasible. Inspired from the core set based approximate MEB algorithms, $(1 + \varepsilon)$ and $(1 - \varepsilon)$-approximation algorithm is utilized for fast training LARM problem, which is called core set based LARM. Firstly, core sets of training patterns are obtained by $(1 + \varepsilon)$ and $(1 - \varepsilon)$-approximation algorithm to achieve the distribution of vector-angular region of the normal and abnormal examples. The core set is a subset of the original training patterns and the optimization problem can be approximately solved on the core set. Secondly, the LARM problem is solved by SMO algorithm [22] using the obtained core set. According to [12, 13], the number of core sets is independent of both the number and the dimension of training patterns, and the time complexity is linear to the number of training patterns while the space complexity is independent of the number of training patterns. The schematic illustration of core set based LARM is shown in Figure 2.

Suppose S_t is the core set of the tth iteration, $\mathbf{v}_t$ is the optimal vector in the feature space of the tth iteration, $\breve{\omega}_t$ is the minimum distance between the center of the vector-angular margin and any point in core set of the tth iteration, and $\widehat{\omega}_t$ is the maximum distance between the center of the vector-angular margin and any point in core set of the tth iteration. Given $\varepsilon > 0$, according to [12, 13], the core set based LARM is trained as follows.

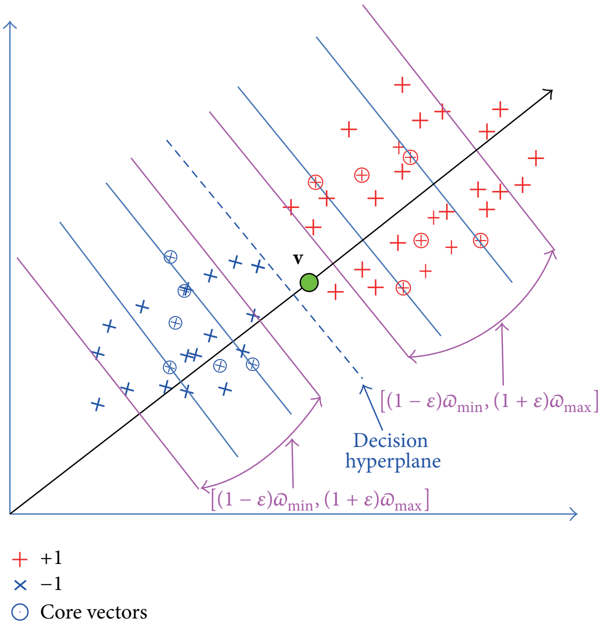

v
$[(1-\varepsilon)\varpi_{\min}, (1+\varepsilon)\varpi_{\max}]$
Decision hyperplane
$[(1-\varepsilon)\varpi_{\min}, (1+\varepsilon)\varpi_{\max}]$
+ +1
× −1
○ Core vectors

Figure 2: Schematic illustration of core set based LARM.

(i) Initialize S_0, ε, $\breve{\varpi}_0$ and $\widehat{\varpi}_0$.

(ii) Terminate if there is no training point $\phi(\mathbf{z})$ falls outside the vector-angular region $[(1-\varepsilon)\times\breve{\varpi}_t, (1+\varepsilon)\times\widehat{\varpi}_t]$. Go to step (vi).

(iii) Find $\mathbf{z}_a$ and $\mathbf{z}_b$; $\mathbf{v}_t{}^{\mathrm{T}}\phi(\mathbf{z}_a)$ is the furthest away from the center of the vector-angular margin and $\mathbf{v}_t{}^{\mathrm{T}}\phi(\mathbf{z}_b)$ is the shortest away from the center of the vector-angular margin. Set $S_{t+1} = S_t \cup \{\mathbf{z}_a, \mathbf{z}_b\}$.

The distance between the center of the vector-angular margin and any point $\phi(\mathbf{z}_\ell)$ is expressed as follows:

$$\begin{aligned}\varpi_t &= \left|\mathbf{v}_t{}^{\mathrm{T}}\phi(\mathbf{z}_\ell) - \omega_t{}^2 + \frac{\rho_t{}^2}{2}\right| \\ &= \left|\sum_{\mathbf{z}_i\in S_t}^{|S_t|}\alpha_i^*\kappa(\mathbf{z}_i,\mathbf{z}_\ell) - \omega_t{}^2 + \frac{\rho_t{}^2}{2}\right|,\end{aligned} \tag{22}$$

where $\omega_t{}^2$ is the width of vector-angular region at the tth iteration, $\rho_t{}^2$ is the vector-angular margin at the tth iteration, and the set $\{\mathbf{z}_\ell\}$ is constructed by all training patterns outside the vector-angular region $[(1-\varepsilon)\times\breve{\varpi}_t, (1+\varepsilon)\times\widehat{\varpi}_t]$.

Computing (22) for all n training patterns, takes $O(|S_t|^2 + n|S_t|) = O(n|S_t|)$ time at the tth iteration. When n is large, time cost will be enormous. In order to reduce the computation cost, the probabilistic speedup method [23] is used to accelerate the vector-angular computations in steps (ii) and (iii). The details of time and space complexities can be seen in [12, 13].

(iv) Find the new vector-angular region $[(1-\varepsilon)\times\breve{\varpi}_{t+1}, (1+\varepsilon)\times\widehat{\varpi}_{t+1}]$.

(v) Increase t by 1 and go back to step (ii).

(vi) Solve the LARM problem (14) by the core set S_t.

(vii) Classify the test pattern by the decision function (19).

4. Experimental Results

The proposed core set based LARM is evaluated on twenty datasets, including both LIBSVM datasets [24] and UCI datasets [25]. Details of the datasets are listed in Table 1, where d is the data dimension, #pos is the total number of normal patterns, #neg is the total number of abnormal patterns, m_1 is the number of normal training patterns, and m_2 is that of abnormal training patterns. The dataset size is ranged from 178 to more than 495,141, and the proportion of major and minor data is ranged from 10 : 1 to 1000 : 1. Experiments are repeated for 10 times with random data partitions, the geometric mean accuracy and the standard deviation are recorded.

4.1. Performance Measurement and Parameter Selection.

The performance of core set based LARM is compared with three kernel-based algorithms: ν-SVM, SVDD, and MAMC. The geometric mean accuracy $g = (a^+ \cdot a^-)^{1/2}$ [26] is used for both parameter selection and algorithm evaluation, where a^+ is the classification accuracy of the positive class and a^- is the classification accuracy of the negative class. The measurement is widely applied in imbalanced data [14, 26, 27], and it considers the classification results on both the positive and the negative classes. To make the experimental results persuasive enough, all the parameters of ν-SVM, SVDD, MAMC, and core set based LARM are selected by fivefold cross validation.

In all experiments, the Radial Basis Function (RBF) is taken as the kernel function:

$$\kappa\left(\mathbf{x}_i,\mathbf{x}_j\right) = \exp\left(-\gamma\left\|\mathbf{x}_i-\mathbf{x}_j\right\|^2\right), \quad 0<\gamma<+\infty, \tag{23}$$

where γ is the kernel parameter of the RBF. For all the algorithms, RBF parameter γ is calculated by [12, 13]

$$\gamma = \frac{n-1}{\sum_{i=1}^n \operatorname{diag}\left(\overline{\mathbf{K}}\right) - (1/n)\sum_{i=1}^n\sum_{j=1}^n \overline{\mathbf{K}}_{i,j}}, \tag{24}$$

where $\overline{\mathbf{K}}_{i,j} = \mathbf{x}_i{}^{\mathrm{T}}\mathbf{x}_j$ and $\operatorname{diag}(\overline{\mathbf{K}})$ is the diagonal elements of matrix $\overline{\mathbf{K}}$.

For ν-SVM, parameter ν is searched in $\{0.1k, 0.01k, 0.001k, 0.0001k\}$, where $k = 1, 3, 5, 7, 9$.

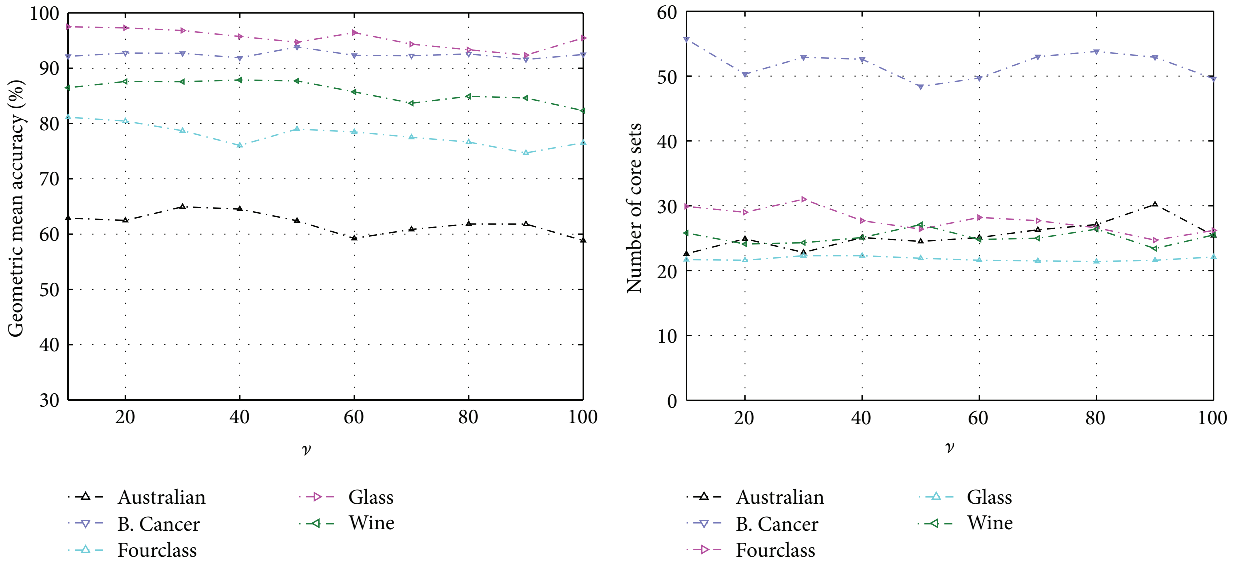
For SVDD, parameter C_1 is searched in $\{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500\}$ and parameter C_2 is searched by the ratio C_2/C_1 belonging to $\{m_1/4m_2, m_1/2m_2, m_1/m_2, 2m_1/m_2, 4m_1/m_2\}$.

For MAMC, parameter C is searched in $\{10, 30, 50, 70, 90\}$ and parameter ν is searched in $\{2, 5, 7, 9, 20, 50, 70, 90\}$.

For core set based LARM, parameter ν is searched in $\{10, 30, 50, 70, 90\}$ and parameters ν_1 and ν_2 are searched in $\{0.001, 0.01\}$. From (21), $\nu_1(\nu+1) \in [0.01, 0.9]$ and $\nu\nu_2 \in [0.01, 0.9]$ can be achieved, which are most associated with the percentage of support vectors and margin errors. From Section 4.2, we can see that parameters λ and ε have faint

Table 1: Datasets used in experiments.

Dataset	#pos	#neg	m_1	m_2	d
Australian	383	307	192	11	14
Banknote authentication (B. authentication)	762	610	381	19	4
Breast Cancer (B. Cancer)	458	241	229	22	9
Cod-rna	39690	19845	19845	99	8
Covtype	283301	211840	141651	141	54
Diabetic	611	540	306	15	19
Fourclass	307	555	154	7	2
Glass	144	70	72	7	11
Heart	120	150	60	6	13
Hill valley with noise (H. valley)	299	307	150	15	100
Ionosphere	225	126	113	5	34
Liver disorders (L. disorders)	200	145	100	10	6
Magic gamma telescope (MC)	12332	6688	6166	12	10
Sensorless drive diagnosis (SDD)	5319	53190	2660	26	48
Skin segmentation (S. segmentation)	194198	50859	97099	97	3
Sonar	111	97	56	5	60
Shuttle	34108	9392	17054	17	9
Svmguide1	2000	2000	1000	10	4
Wilt	4265	74	2133	10	5
Wine	119	59	84	8	13



(a) Influence of ν on geometric mean accuracy

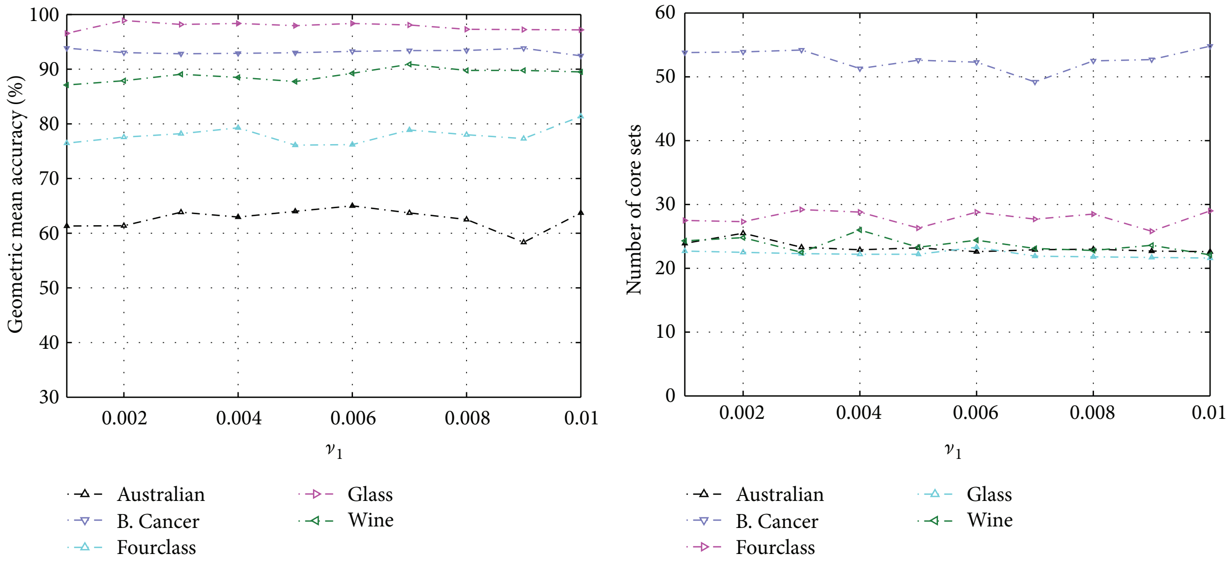
(b) Influence of ν on the number of core sets

Figure 3: Influence of parameter ν on geometric mean accuracy and the number of core sets.

effect on the accuracy rate. Therefore, parameters λ and ε are set to 1 and 10^{-5}, respectively.

4.2. Parameters Influence. There are five parameters in core set based LARM, that is, ν, ν_1, ν_2, λ, and ε. To verify the influence of the parameters on the performance of core set based LARM, experiments on some representative datasets are performed. By fixing other parameters, the influence of every parameter on some representative datasets is further studied, which is shown in Figures 3–7.

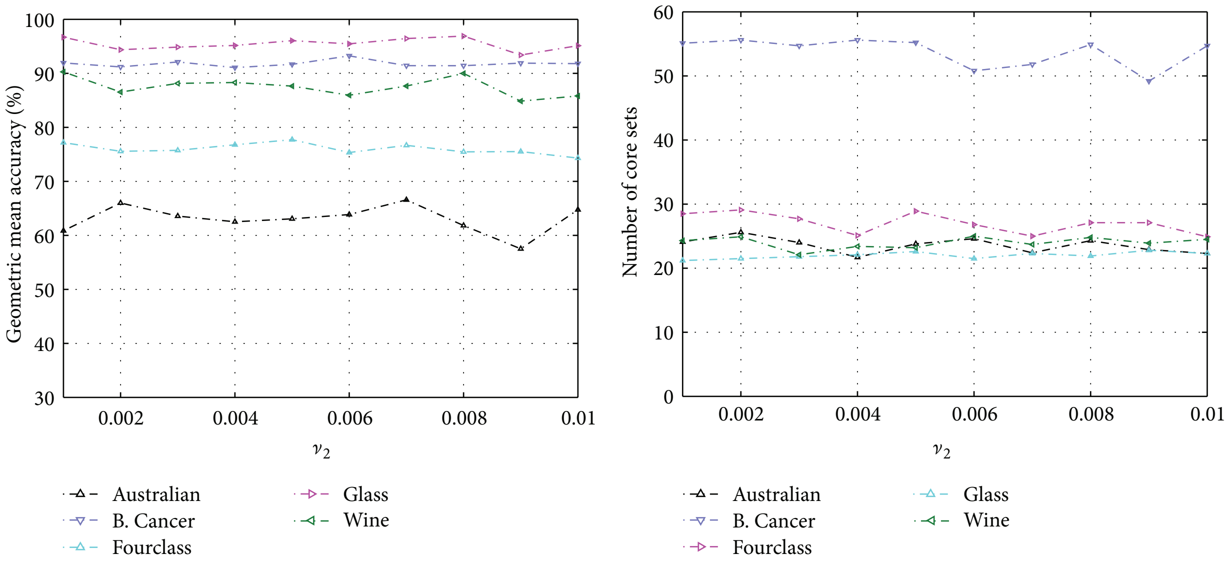
Figure 3 shows the influence of ν on the geometric mean accuracy and the number of core sets by varying ν from 10 to 100 while fixing ν_1, ν_2, λ, and ε as the suggested value obtained by the cross validation described in Section 4.1. Figure 4 shows the influence of ν_1 on the geometric mean accuracy and the number of core sets by varying ν_1 from 0.001 to 0.01 while fixing ν, ν_2, λ, and ε in the same way. Figure 5 shows the influence of ν_2 on the geometric mean accuracy and the number of core sets by varying ν_2 from 0.001 to 0.01 while fixing ν, ν_1, λ, and ε in the same way. Figure 6 shows



(a) Influence of ν_1 on geometric mean accuracy

(b) Influence of ν_1 on the number of core sets

FIGURE 4: Influence of parameter ν_1 on geometric mean accuracy and the number of core sets.



(a) Influence of ν_2 on geometric mean accuracy

(b) Influence of ν_2 on the number of core sets

FIGURE 5: Influence of parameter ν_2 on geometric mean accuracy and the number of core sets.

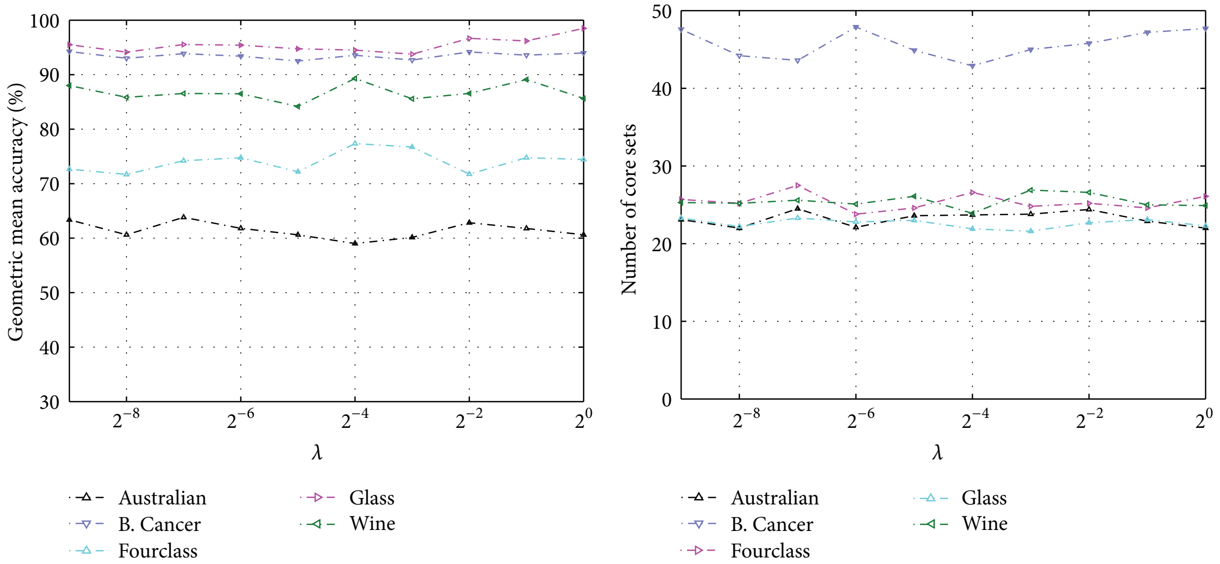
the influence of λ on the geometric mean accuracy and the number of core sets by varying λ from 2^{-9} to 2^0 while fixing ν, ν_1, ν_2, and ε in the same way. Figure 7 shows the influence of ε on the geometric mean accuracy and the number of core sets by varying ε from $1e-9$ to $1e-2$ while fixing ν, ν_1, ν_2, and λ in the same way.

From Figures 3–7, it can be seen that parameters ν, ν_1, ν_2, λ, and ε have faint effect on the geometric mean accuracy and the number of core sets, which make the core set based LARM even more attractive in practice. Therefore, parameters ν, ν_1, ν_2, λ, and ε obtained by the cross validation described in Section 4.1 are acceptable for all experiments.

4.3. Numerical Results

4.3.1. Detection Performance. For each dataset, samples are randomly split into training patterns and testing patterns with the proportion described in Table 1. Parameters of ν-SVM, SVDD, MAMC, and core set based LARM are selected by fivefold cross validation to make the experimental results persuasive enough.

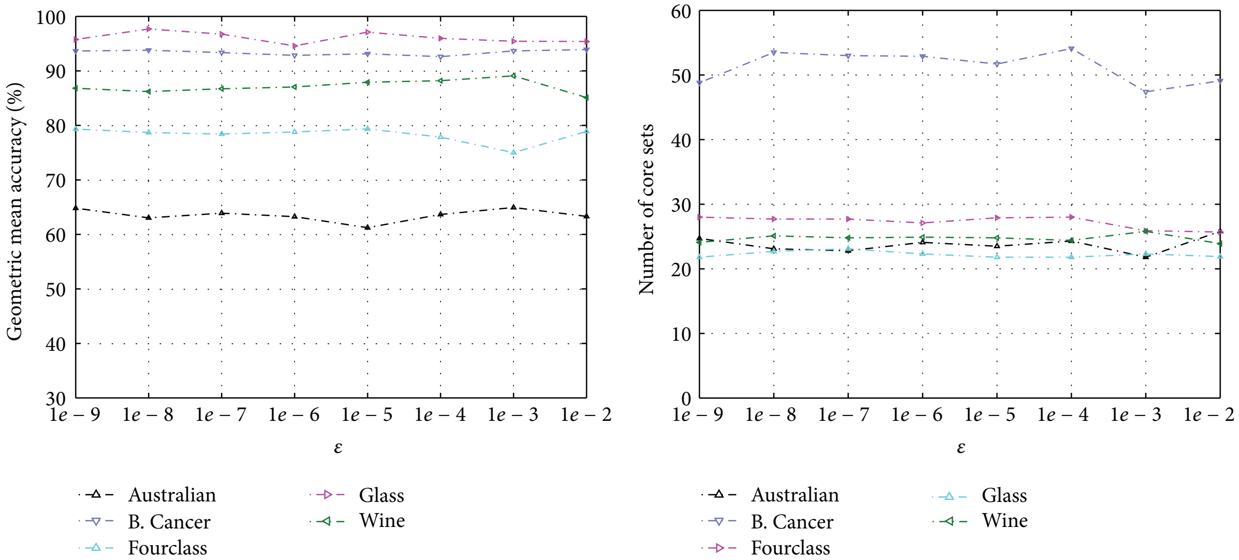
The geometric mean accuracy is used for the performance evaluation. Experiments are repeated for 10 times with random data partitions. The average accuracy and the standard deviation are listed in Table 2. NULL shows that



(a) Influence of λ on geometric mean accuracy

(b) Influence of λ on the number of core sets

Figure 6: Influence of parameter λ on geometric mean accuracy and the number of core sets.



(a) Influence of ε on geometric mean accuracy

(b) Influence of ε on the number of core sets

Figure 7: Influence of parameter ε on geometric mean accuracy and the number of core sets.

there is no return result in 10 hours. Furthermore, with regard to every dataset, the difference between the bold results and the best geometric mean accuracy is not significant, which is determined by the Wilcoxon rank-sum test, with the confidence level of 0.05.

From Table 2, it can be concluded that the performance of core set based LARM is comparable to the best of ν-SVM, SVDD, and MAMC on all datasets. The core set based LARM performs significantly better than ν-SVM, SVDD, and MAMC on 12, 9, and 13 over 20 datasets, respectively. It illustrates that, by using $(1+\varepsilon)$ and $(1-\varepsilon)$-approximation algorithm for training LARM, the generalization performance of core set based LARM is comparable to or even better than the best of ν-SVM, SVDD, and MAMC.

4.3.2. Time Cost. The time cost of ν-SVM, SVDD, MAMC, and core set based LARM on different datasets is shown in Tables 3 and 4. The average and standard deviation of training time (including parameters selection and model training time) are shown in Table 3. The average and standard deviation of testing time are shown in Table 4. All the experiments are conducted on the computer with an i5-2400@3.10 GHz CPU and 8 GB SDRAM. NULL shows that there is no return result in 10 hours. Furthermore, with regard

Table 2: Average geometric mean accuracy and standard deviation on datasets.

Dataset	ν-SVM (%)	SVDD (%)	MAMC (%)	Core set based LARM (%)
Australian	47.88 ± 14.46	56.03 ± 9.21	35.11 ± 10.36	**63.67 ± 9.12**
B. authentication	**97.25 ± 1.66**	**98.60 ± 0.95**	95.34 ± 3.65	**98.60 ± 2.30**
B. Cancer	92.38 ± 1.43	**95.25 ± 0.89**	92.75 ± 2.43	**93.61 ± 3.53**
Cod-rna	45.63 ± 15.18	NULL	NULL	**75.46 ± 8.79**
Covtype	NULL	NULL	NULL	**57.51 ± 8.27**
Diabetic	42.91 ± 9.34	**52.26 ± 9.01**	**50.14 ± 8.54**	**53.50 ± 8.90**
Fourclass	**77.23 ± 7.34**	**81.74 ± 11.27**	64.31 ± 13.04	**81.97 ± 8.16**
Glass	**95.12 ± 4.27**	80.78 ± 5.71	**94.56 ± 4.47**	**96.56 ± 3.52**
Heart	42.01 ± 14.10	**52.23 ± 5.80**	**51.50 ± 10.84**	**54.29 ± 8.32**
H. valley	**43.05 ± 9.25**	**42.00 ± 10.79**	**35.76 ± 19.45**	**44.16 ± 2.34**
Ionosphere	46.44 ± 16.10	**66.54 ± 9.81**	36.07 ± 16.80	**71.17 ± 14.91**
L. disorders	50.21 ± 5.83	**54.72 ± 7.49**	45.80 ± 6.58	**58.55 ± 3.14**
MC	20.46 ± 3.84	**67.36 ± 3.87**	41.57 ± 7.12	**62.94 ± 5.46**
SDD	**40.26 ± 9.63**	25.93 ± 9.54	20.64 ± 21.66	**45.67 ± 10.30**
S. segmentation	NULL	NULL	NULL	**95.91 ± 1.92**
Sonar	**52.27 ± 8.87**	31.39 ± 6.99	**55.79 ± 11.36**	**46.37 ± 9.04**
Shuttle	**91.65 ± 5.49**	40.23 ± 21.59	NULL	**92.88 ± 4.11**
Svmguide1	82.17 ± 6.91	**90.21 ± 6.88**	**89.95 ± 4.18**	**91.83 ± 2.69**
Wilt	**82.14 ± 9.29**	56.95 ± 15.52	64.76 ± 3.81	**84.95 ± 13.94**
Wine	82.32 ± 4.57	**86.00 ± 6.95**	**87.19 ± 4.32**	**87.62 ± 1.97**

Table 3: Training time on different datasets.

Dataset	ν-SVM (s)	SVDD (s)	MAMC (s)	Core set based LARM (s)
Australian	**0.0739 ± 0.0086**	0.5975 ± 0.0993	0.6325 ± 0.0122	0.8819 ± 0.4613
B. authentication	**0.1627 ± 0.0699**	0.3441 ± 0.0199	1.3943 ± 0.0131	1.1103 ± 0.1658
B. Cancer	**0.0975 ± 0.0063**	0.2462 ± 0.0152	0.7309 ± 0.0126	2.2633 ± 0.2803
Cod-rna	388.7901 ± 24.6651	NULL	NULL	**7.4578 ± 0.1622**
Covtype	NULL	NULL	NULL	**64.9519 ± 7.3727**
Diabetic	**0.2096 ± 0.0063**	1.1188 ± 0.0819	1.9243 ± 0.0119	5.7415 ± 1.0748
Fourclass	**0.0354 ± 0.0030**	0.1271 ± 0.0151	0.2706 ± 0.0117	0.5652 ± 0.1175
Glass	**0.0255 ± 0.0038**	0.0587 ± 0.0038	0.1459 ± 0.0088	0.2079 ± 0.0220
Heart	**0.0229 ± 0.0016**	0.0858 ± 0.0082	0.1148 ± 0.0070	0.8305 ± 0.2195
H. valley	**0.1457 ± 0.0119**	1.0385 ± 0.0278	1.4355 ± 0.0164	0.8506 ± 0.2926
Ionosphere	**0.0634 ± 0.0097**	0.1999 ± 0.0134	0.4018 ± 0.0087	1.9758 ± 0.3947
L. disorders	**0.0399 ± 0.0058**	0.1861 ± 0.0202	0.1809 ± 0.0095	1.6280 ± 0.4039
MC	19.5417 ± 0.3331	49.0111 ± 10.8097	472.2837 ± 1.3679	**17.1713 ± 5.3223**
SDD	13.8246 ± 1.0316	145.7426 ± 18.0463	218.5634 ± 0.6386	**8.3123 ± 0.5485**
S. segmentation	NULL	NULL	NULL	**7.2536 ± 0.3645**
Sonar	**0.0463 ± 0.0050**	0.1406 ± 0.0061	0.1896 ± 0.0054	2.3453 ± 0.4481
Shuttle	226.0020 ± 3.1471	268.0641 ± 145.7989	NULL	**7.3061 ± 0.2336**
Svmguide1	**0.4634 ± 0.0171**	1.3081 ± 0.3742	8.9181 ± 0.0539	1.0625 ± 0.0781
Wilt	**2.0750 ± 0.0774**	8.6527 ± 0.7402	52.8582 ± 0.7763	3.6379 ± 0.3202
Wine	**0.0263 ± 0.0049**	0.0983 ± 0.0274	0.1785 ± 0.0111	0.2936 ± 0.0494

to every dataset, the difference between the bold results and the best time cost is not significant, which is determined by the Wilcoxon rank-sum test, with the confidence level of 0.05.

From Table 3, it can be clearly seen that the training time of core set based LARM is longer than the best of ν-SVM, SVDD, and MAMC, when the number of the training patterns is less than 2,143. However, when the number of training patterns is larger than 2,686 such as SDD, MC, Shuttle, Cod-rna, S. segmentation, and Covtype, the training time of core set based LARM is shorter than the best of ν-SVM, SVDD, and MAMC. When the number of training patterns increases to 141,792, the average training time of core

Table 4: Testing time on different datasets.

Dataset	ν-SVM (s)	SVDD (s)	MAMC (s)	Core set based LARM (s)
Australian	**0.0009 ± 0.0005**	0.0035 ± 0.0014	0.0023 ± 0.0017	**0.0013 ± 0.0004**
B. authentication	**0.0050 ± 0.0129**	**0.0010 ± 0.0004**	0.0070 ± 0.0032	0.0018 ± 0.0012
B. Cancer	**0.0011 ± 0.0003**	**0.0013 ± 0.0004**	0.0032 ± 0.0012	**0.0012 ± 0.0003**
Cod-rna	0.2901 ± 0.2182	NULL	NULL	**0.0584 ± 0.0085**
Covtype	NULL	NULL	NULL	**1.5096 ± 1.1866**
Diabetic	**0.0015 ± 0.0009**	0.0054 ± 0.0016	0.0075 ± 0.0056	0.0029 ± 0.0010
Fourclass	**0.0004 ± 0.0001**	0.0006 ± 0.0002	0.0015 ± 0.0007	0.0007 ± 0.0002
Glass	**0.0002 ± 0.0002**	**0.0002 ± 0.0001**	0.0005 ± 0.0004	0.0004 ± 0.0003
Heart	**0.0003 ± 0.0001**	0.0005 ± 0.0001	0.0008 ± 0.0004	0.0007 ± 0.0004
H. valley	**0.0021 ± 0.0013**	0.0101 ± 0.0021	0.0050 ± 0.0032	**0.0031 ± 0.0009**
Ionosphere	**0.0007 ± 0.0003**	0.0016 ± 0.0008	0.0018 ± 0.0009	0.0011 ± 0.0003
L. disorders	**0.0004 ± 0.0002**	0.0006 ± 0.0002	0.0006 ± 0.0002	0.0006 ± 0.0003
MC	**0.0282 ± 0.0073**	0.6090 ± 0.3269	0.7920 ± 0.9542	**0.0269 ± 0.0132**
SDD	**0.2947 ± 0.3511**	9.4060 ± 3.3205	2.6391 ± 4.2971	**0.1829 ± 0.0488**
S. segmentation	NULL	NULL	NULL	**0.1504 ± 0.0331**
Sonar	**0.0006 ± 0.0003**	0.0011 ± 0.0003	0.0013 ± 0.0003	0.0010 ± 0.0004
Shuttle	**0.0393 ± 0.0147**	4.5419 ± 3.2784	NULL	**0.0342 ± 0.0082**
Svmguide1	**0.0027 ± 0.0009**	**0.0031 ± 0.0021**	0.0696 ± 0.0214	0.0039 ± 0.0010
Wilt	**0.0020 ± 0.0003**	0.0052 ± 0.0056	0.1133 ± 0.0230	0.0040 ± 0.0009
Wine	**0.0002 ± 0.0004**	0.0003 ± 0.0003	0.0004 ± 0.0003	0.0003 ± 0.0001

set based LARM does not exceed 65 seconds. Therefore, the training time of core set based LARM does not increase very quickly with the number of training patterns.

As can be seen from Table 4, the best testing time of ν-SVM, SVDD, and MAMC performs slightly better than core set based LARM on 11 over 20 datasets; the longest time gap is 0.002 second. However, the testing time of core set based LARM is not the worst one. When the number of testing patterns is 353,349, such as Covtype, the average testing time of core set based LARM is about 1.5 seconds. It shows that the core set based LARM can detect testing examples fast.

5. Conclusion

In this paper, a novel LARM algorithm and its fast training method based on core set are proposed for novelty detection on imbalanced data. The proposed LARM algorithm combines the ideas of one-class and binary classification algorithms, which constructs the largest vector-angular region in the feature space to separate normal training patterns and maximizes the vector-angular margin between this optimal vector-angular region and the abnormal data. In order to make the generalization performance of LARM better, the vector-angular distribution is optimized by maximizing the vector-angular mean and minimizing the vector-angular variance. To improve the computation efficiency, $(1+\varepsilon)$ and $(1-\varepsilon)$-approximation algorithm is proposed for fast training LARM based on core set. The time and space complexity of core set based LARM are linear to and independent of the number of training patterns, respectively. Comprehensive experiments have validated the effectiveness of proposed approach. In the future, it will be interesting to extend the idea of LARM to handle one-class learning problem.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Natural Science Foundation of China under Grant nos. U1433103 and U1333116; the Science and Technology Foundation of Civil Aviation Administration of China under Grant no. 20150227; and the Fundamental Research Foundation for the Central Universities of CAUC under Grant no. 3122014D022 and no. 3122014B002.

References

[1] R. Perdisci, G. Gu, and W. Lee, "Using an ensemble of one-class SVM classifiers to harden payload-based anomaly detection systems," in *Proceedings of the 6th International Conference on Data Mining (ICDM '06)*, pp. 488–498, Hong Kong, December 2006.

[2] P. Hayton, S. Utete, D. King, S. King, P. Anuzis, and L. Tarassenko, "Static and dynamic novelty detection methods for jet engine health monitoring," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 365, no. 1851, pp. 493–514, 2007.

[3] L. Clifton, D. A. Clifton, P. J. Watkinson, and L. Tarassenko, "Identification of patient deterioration in vital-sign data using one-class support vector machines," in *Proceedings of the*

Federated Conference on Computer Science and Information Systems (FedCSIS '11), pp. 125–131, September 2011.

[4] E. Smart and D. Brown, "A two-phase method of detecting abnormalities in aircraft flight data and ranking their impact on individual flights," *IEEE Transactions on Intelligent Transportation Systems*, vol. 13, no. 3, pp. 1253–1265, 2012.

[5] R. X. Guo, K. Guo, and J. K. Dong, "Fault diagnosis for the landing phase of the aircraft based on an adaptive kernel principal component analysis algorithm," *Proceedings of the Institution of Mechanical Engineers Part I: Journal of Systems & Control Engineering*, vol. 229, no. 10, pp. 917–926, 2015.

[6] B. Schölkopf, R. Williamson, A. Smola, J. Shawe-Taylor, and J. Platt, "Support vector method for novelty detection," in *Advances in Neural Information Processing Systems—NIPS 1999*, pp. 582–588, MIT Press, 1999.

[7] D. M. J. Tax and R. P. W. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, 2004.

[8] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *The Journal of Machine Learning Research*, vol. 2, no. 2, pp. 139–154, 2002.

[9] M. Wu and J. Ye, "A small sphere and large margin approach for novelty detection using training data with outliers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 2088–2092, 2009.

[10] B. Krawczyk and M. Woźniak, "Incremental weighted one-class classifier for mining stationary data streams," *Journal of Computational Science*, vol. 9, pp. 19–25, 2015.

[11] B. Krawczyk and M. Woźniak, "One-class classifiers with incremental learning and forgetting for data streams with concept drift," *Soft Computing*, vol. 19, no. 12, pp. 3387–3400, 2015.

[12] I. W. Tsang, J. T. Kwok, and P.-M. Cheung, "Core vector machines: fast SVM training on very large data sets," *Journal of Machine Learning Research*, vol. 6, pp. 363–392, 2005.

[13] I. W. H. Tsang, J. T. Y. Kwok, and J. A. Zurada, "Generalized core vector machines," *IEEE Transactions on Neural Networks*, vol. 17, no. 5, pp. 1126–1140, 2006.

[14] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[15] W. Hu, F.-L. Chung, and S. Wang, "The maximum vector-angular margin classifier and its fast training on large datasets using a core vector machine," *Neural Networks*, vol. 27, no. 3, pp. 60–73, 2012.

[16] M. A. F. Pimentel, D. A. Clifton, L. Clifton, and L. Tarassenko, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014.

[17] W. Gao and Z.-H. Zhou, "On the doubt about margin explanation of boosting," *Artificial Intelligence*, vol. 203, pp. 1–18, 2013.

[18] T. Zhang and Z.-H. Zhou, "Large margin distribution machine," in *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '14)*, pp. 313–322, New York, NY, USA, August 2014.

[19] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, "New support vector algorithms," *Neural Computation*, vol. 12, no. 5, pp. 1207–1245, 2000.

[20] C.-C. Chang and C.-J. Lin, "Training ν-support vector classifiers: theory and algorithms," *Neural Computation*, vol. 13, no. 9, pp. 2119–2147, 2001.

[21] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[22] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods—Support Vector Learning*, B. Schölkopf, C. Burges, and A. Smola, Eds., pp. 185–208, MIT Press, Cambridge, Mass, USA, 1999.

[23] A. Smola and B. Schölkopf, "Sparse greedy matrix approximation for machine learning," in *Proceedings of the 17th International Conference on Machine Learning (ICML '00)*, pp. 911–918, Stanford, Calif, USA, June 2000.

[24] R. E. Fan and C. J. Lin, *LIBSVM Data: Classification, Regression and Multi-Label*, 2011, https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets.

[25] A. Asuncion and D. J. Newman, *UCI Machine Learning Repository*, School of Information and Computer Sciences, University of California Irvine, 2007, http://www.ics.uci.edu/~mlearn/MLRepository.html.

[26] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: one-sided selection," in *Proceedings of the 14th International Conference on Machine Learning*, pp. 179–186, Morgan Kaufmann, 1997.

[27] G. Wu and E. Y. Chang, "Class-boundary alignment for imbalanced dataset learning," in *Proceedings of the International Conference on Machine Learning Workshop Learning from Imbalanced Datasets (ICML '03)*, 2003.

Advances in
Operations Research
Advances in
Decision Sciences
Journal of
Applied Mathematics
Algebra
Journal of
Probability and Statistics
The Scientific
World Journal
International Journal of
Differential Equations
Hindawi
Submit your manuscripts at
http://www.hindawi.com
International Journal of
Combinatorics
Advances in
Mathematical Physics
Journal of
Complex Analysis
Journal of
Mathematics
Mathematical Problems
in Engineering
Abstract and
Applied Analysis
Discrete Dynamics in
Nature and Society
International
Journal of
Mathematics and
Mathematical
Sciences
Journal of
Discrete Mathematics
Journal of
Function Spaces
International Journal of
Stochastic Analysis
Journal of
Optimization