AMSE JOURNALS-2016-Series: Advances B; Vol. 59; N°1; pp 46-61
Submitted July 2016; Revised Oct. 15, 2016, Accepted Dec. 10, 2016

Multiple Kernel Sphere with Large Margin for Novelty Detection

X. M. Cheng, B. Zheng, W. J. Hu

School of Information Engineering, Huzhou University
Huzhou 313000, Zhejiang, China (chengxinmin@hutc.zj.cn, zb83dana@163.com, hoowenjun@foxmail.com)

Abstract

Novelty detection methods have been frequently applied in medical diagnosis, fault detection, network security and the discovery of new species. Among them, Support Vector Data Description (SVDD) has received considerable attention for its comprehensivedescription ability which covers the target data. Additionally, the Multiple Kernel Learning (MKL) technique has been extensively applied in machine learning methods; e.g. the SVM classifiers, dimensionality reduction techniques, etc. In this paper, we focus on the application of the MKL method on novelty detection (ND) and propose the new method of Multiple Kernel Sphere with Larger Margin (MKSLM) for novelty detection. In the presented method, the volume of the sphere is minimized while the margin between the surface of the sphere and the outliers are maximized to obtain a sphere with minimum size. An algorithm is also developed to solve the optimization problem. Experimental results over various real data sets have validated the superiority of the proposed methods

Key words

Novelty detection, multiple kernel, support vector data description, classification margin.

1. Introduction

Novelty detection (ND) is a method to identify the outliers that are different from the ordinary data. The outliers are also called abnormal samples or novelty samples. In practice, the amount of normal data is much more than that of abnormal data, which makes it difficult and computationally expensive to identify the abnormal data. Therefore, most of the ND methods focus only on the processing of normal data; e.g., One-Class Support Vector Data Description

(SVDD)1 and One-Class Support Vector Machine (OCSVM)2. As a result, ND is often named a One-Class Classification (OCC) problem. The ND approaches are usually applied in currency validation, medical diagnosis 3, fault detection 45, network security6, new species discovery7, etc.

ND approaches have been investigated extensively in the past several decades and a number of techniques have been developed, including neural network-based methods 8 9, density estimation-based methods 10 and spatial depth-based methods 7. Recently, Wu et al. proposed SSLM approach 11. This method takes the classification margins of the binary classification algorithms into account; e.g., $2/\|\mathbf{w}\|$ for Support Vector Machine (SVM) 12and $\rho/\|\mathbf{w}\|$ for v-SVM13, where ρ and $\mathbf{W}$ are the optimal margin (note that it is not the classification margin) and the normal vector of the hyperplane respectively. Additionally, Cha et al. introduced the density weight to SVDD for searching an optimal description, in which the weight is estimated through the relative density of each data point with k-NN approach, so that the optimal description prioritizes the data points in high-density regions and eventually shifts to these regions 14. Wang et al. assigned a position-based weight parameter to each data point, which is computed according to the distance between the corresponding mapped point in the feature space and the mean of feature space, and used the weight parameters to replace the SVDD's trade-off parameter for overcoming the sensitivity to the selection of the trade-off parameter 15. Schölkopf et al. proposed the OCSVM approach for the One-Class Classification (OCC) problem 2 by extending the large margin of SVM 12.

Furthermore, researchers have also studied the Multiple Kernel Learning (MKL) technique in machine learning 161718. For example, MKL is employed and embedded in the training process of SVM 1920, dimensionality reduction techniques21, domain transfer classifiers22, and other applications232425. Recently, it has been introduced to the graph regularized NMF 26. Generally, SVDD may obtain the optimal description only with a kernel and a suitable parameter for handling the simple nonlinear distribution of data samples. However, it still cannot capture the multi distribution of the target data. Furthermore, it is unclear whether MKL can be applied to OCC problems, especially for SVDD. Therefore, we have investigated this issue here and propose the Multiple Kernel Sphere with Larger Margin (MKSLM) approach for ND.

2. Related work

The ND methods most related to our approach include SVDD and SSLM. A brief description of these two methods is provided in this section. Let $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N]$ be the data matrix with $\mathbf{x}_i \in \Re^D$, and $\mathbf{y} = [y_1, \cdots, y_N]^T$ be the class label vector, with $y_i = 1$ if $\mathbf{x}_i$ being a normal sample; and $y_i = -1$ otherwise. Without a loss of generality, assume $\mathbf{x}_i (1 \le i \le m_1)$ is a normal sample, and $\mathbf{x}_j$ $(m_1 + 1 \le j \le m_1 + m_2 = N)$ is an abnormal sample or an outlier.

2.1 SVDD

SVDD models can be divided into two categories 1. One is called the one-class SVDD, which is constructed only with the normal data. The other is called the two-class SVDD, which is constructed with both the normal data and the abnormal data. Both of them aim to construct a small sphere with center $\mathbf{c}$ and radius r, so that the volume is as small as possible. The two-class SVDD can be formulated as:

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

where $C_1 \ge 0$ and $C_2 \ge 0$ are two parameters used to control the tradeoff between the sphere volume and the error, and $\xi_i \ge 0 (1 \le i \le N)$ is the slack variable. By using the Lagrangian technique, we can obtain the dual problem:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \ \sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j), \\ &\text{s.t.} \ \sum_{i=1}^{N} \alpha_i y_i = 1, \ \ 0 \le \alpha_i \le C_1, 1 \le i \le m_1, \ \ 0 \le \alpha_j \le C_2, m_1 + 1 \le i \le N, \end{aligned}$$

where $\boldsymbol{\alpha} = [\alpha_1, \cdots, \alpha_N]^T$ is the Lagrangian multiplier vector, and $k(\bullet, \bullet)$ is a kernel function. Following some simple algebraic steps, the above dual problem can be rewritten as follows:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \ (\boldsymbol{\alpha} \circ \mathbf{y})^T diag(\mathbf{K}) - (\boldsymbol{\alpha} \circ \mathbf{y})^T \mathbf{K} (\boldsymbol{\alpha} \circ \mathbf{y}), \\ &\text{s.t.} \ \mathbf{1}^T (\boldsymbol{\alpha} \circ \mathbf{y}) = 1, \ \ 0.5 C_2 (\mathbf{y} - \mathbf{1}) \le (\boldsymbol{\alpha} \circ \mathbf{y}) \le 0.5 C_1 (\mathbf{y} + \mathbf{1}), \end{aligned} \tag{2}$$

where $\mathbf{1}$ is a vector of ones, $\mathbf{K}=[\phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)]_{N\times N}=[k(\mathbf{x}_i,\mathbf{x}_j)]_{N\times N}$ is the corresponding kernel matrix associated with the kernel function $k(\bullet,\bullet)$, and the sign "$\circ$" is the Hadamard product operator.

After obtaining the dual variables $\boldsymbol{\alpha}$, an unknown data sample $\mathbf{x}$ can be classified by using the following decision function:

$$f_{\mathrm{SVDD}}(\mathbf{x}) = r^2 - \left\|\phi(\mathbf{x}) - \mathbf{c}\right\|^2 = r^2 - k(\mathbf{x},\mathbf{x}) + 2\sum_{i=1}^{N}\alpha_i y_i k(\mathbf{x}_i,\mathbf{x}) - \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j), \tag{3}$$

where $r^2=\frac{1}{|S_0|}\sum_{\mathbf{x}_i\in S_0}\left\|\phi(\mathbf{x}_i)-\mathbf{c}\right\|^2$ with $S_0=\{\mathbf{x}_i\,|\,0<\alpha_i<C_1,1\le i\le m_1\}\cup\{\mathbf{x}_i\,|\,0<\alpha_j<C_2,m_1+1\le j\le N\}$.

$\mathbf{x}$ is a normal sample if $f(\mathbf{x})\ge 0$. Otherwise, it is an outlier or an abnormal sample.

2.2 SSLM

SSLM is proposed for ND to deal with situations where the training data contain many normal samples and few outliers; i.e., abnormal samples11. The method is advantageous in that it takes into consideration the margin between the outliers and the surface of the sphere while constructing this sphere. The SSLM model is formulated as

$$\begin{aligned}
&\min_{r,\mathbf{c}}\ r^2-\mu\rho^2+\frac{1}{v_1m_1}\sum_{i=1}^{m_1}\xi_i+\frac{1}{v_2m_2}\sum_{j=m_1+1}^{N}\xi_j,\\
&\text{s.t. }\left\|\phi(\mathbf{x}_i)-\mathbf{c}\right\|^2\le r^2+\xi_i,1\le i\le m_1,\\
&\quad\left\|\phi(\mathbf{x}_j)-\mathbf{c}\right\|^2\ge r^2+\rho^2+\xi_j,m_1+1\le j\le N,\ \xi_i\ge 0,1\le i\le N,
\end{aligned} \tag{4}$$

where $\mu\ge 0$, $v_1\ge 0$ and $v_2\ge 0$ are three parameters used to control the tradeoff among the sphere volume, the margin and the error, ρ^2 is the margin between the outliers and the surface of the sphere, and $\xi_i\ge 0\,(1\le i\le N)$ is the slack variable. By using the Lagrangian technique, the dual problem of (4) is given as:

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}}\ (\boldsymbol{\alpha}\circ\mathbf{y})^T diag(\mathbf{K})-(\boldsymbol{\alpha}\circ\mathbf{y})^T\mathbf{K}(\boldsymbol{\alpha}\circ\mathbf{y}),\\
&\text{s.t. }\mathbf{1}^T(\boldsymbol{\alpha}\circ\mathbf{y})=1,\ \frac{0.5}{v_2m_2}(\mathbf{y}-\mathbf{1})\le(\boldsymbol{\alpha}\circ\mathbf{y})\le\frac{0.5}{v_1m_1}(\mathbf{y}+\mathbf{1}),\\
&\quad\mathbf{y}^T(\boldsymbol{\alpha}\circ\mathbf{y})=2\mu+1.
\end{aligned} \tag{5}$$

From the above QP problem, we can obtain the dual variables $\boldsymbol{\alpha}$. Then, an unknown data sample $\mathbf{x}$ can be detected by using the following decision function:

$$f_{\mathrm{SSLM}}(\mathbf{x}) = r^2 - \left\|\phi(\mathbf{x}) - \mathbf{c}\right\|^2 = r^2 - k(\mathbf{x},\mathbf{x}) + 2\sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}_i,\mathbf{x}) - \sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i,\mathbf{x}_j), \tag{6}$$

where $r^2 = \frac{1}{|S_1|}\sum_{\mathbf{x}_i \in S_1}\left\|\phi(\mathbf{x}_i) - \mathbf{c}\right\|^2$, $S_1 = \{\mathbf{x}_i \mid 0 < \alpha_i < \frac{1}{v_1 m_1}, 1 \le i \le m_1\}$. $\mathbf{x}$ is a normal data sample if $f(\mathbf{x}) \ge 0$. Otherwise, it is an outlier, or an abnormal data sample.

3. Multiple kernel sphere with large margin

In this section, a new approach is proposed for ND, that is Multiple Kernel Sphere with Larger Margin (MKSLM). In addition, a technique similar to the Simple Multiple Kernel Learning (SimpleMKL) algorithm is also introduced to solve the optimization problem in MKSLM19.

3.1 MKSLM Primal Problem

Assume the kernel function $k(\bullet,\bullet)$ (k in short) is a linear combination of a set of base kernels; i.e., $k(\mathbf{x}_i,\mathbf{x}_j) = \sum_{m=1}^{M} d_m k_m(\mathbf{x}_i,\mathbf{x}_j)$, where $k_m(\mathbf{x}_i,\mathbf{x}_j) = \phi_m(\mathbf{x}_i)^T \phi_m(\mathbf{x}_j)$, $d_m \ge 0$ is the kernel coefficient and $\sum_{m=1}^{M} d_m = 1$. Then, the kernel matrix of each base kernel is $\mathbf{K}_m = [k_m(\mathbf{x}_i,\mathbf{x}_j)]_{N\times N}$ and the corresponding kernel matrix $\mathbf{K} = \sum_{m=1}^{M} d_m \mathbf{K}_m$. To better describe the MKSLM, the theorem below is proposed.

Theorem 1. Given multiple base kernels and the kernel combination coefficient $\mathbf{d} = [d_1,\cdots,d_m]^T$, the sphere constructed by SSLM in (4) is given as:

$$\begin{aligned} &R(r,\mathbf{c}_m) = \min_{r,\mathbf{c}}\ r^2 - \mu\sum_{m=1}^{M} d_m \rho_m^2 + \frac{1}{v_1 m_1}\sum_{i=1}^{m_1}\xi_i + \frac{1}{v_2 m_2}\sum_{j=m_1+1}^{N}\xi_j, \\ &\text{s.t. } \sum_{m=1}^{M} d_m \left\|\phi_m(\mathbf{x}_i) - \mathbf{c}_m\right\|^2 \le r^2 + \xi_i, 1 \le i \le m_1, \\ &\quad \sum_{m=1}^{M} d_m \left\|\phi_m(\mathbf{x}_j) - \mathbf{c}_m\right\|^2 \ge r^2 + \sum_{m=1}^{M} d_m \rho_m^2 - \xi_j, \\ &\quad m_1 + 1 \le j \le N,\ \xi_i \ge 0, 1 \le i \le N. \end{aligned} \tag{7}$$

Proof: Given a set of base kernels $k_m(\mathbf{x}_i, \mathbf{x}_j) = \phi_m(\mathbf{x}_i)^T \phi_m(\mathbf{x}_j)$ and define $\mathbf{c}_m = \sum_{i=1}^{N} \alpha_i y_i \phi_m(\mathbf{x}_i)$.

Since $\mathbf{c} = \sum_{i=1}^{N} \alpha_i y_i \phi(\mathbf{x}_i)$ in SSLM, then

$$\begin{aligned} \mathbf{c}^T \phi(\mathbf{x}) &= \sum_{i=1}^{N} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) = \sum_{i=1}^{N} \alpha_i y_i \sum_{m=1}^{M} d_m k_m(\mathbf{x}_i, \mathbf{x}) = \sum_{m=1}^{M} d_m \sum_{i=1}^{N} \alpha_i y_i k_m(\mathbf{x}_i, \mathbf{x}) \\ &= \sum_{m=1}^{M} d_m \sum_{i=1}^{N} \alpha_i y_i \phi_m(\mathbf{x}_i)^T \phi_m(\mathbf{x}) = \sum_{m=1}^{M} d_m \mathbf{c}_m^T \phi_m(\mathbf{x}) \end{aligned}, \tag{8}$$

and

$$\begin{aligned} \mathbf{c}^T \mathbf{c} &= \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) = \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \sum_{m=1}^{M} d_m k_m(\mathbf{x}_i, \mathbf{x}_j) \\ &= \sum_{m=1}^{M} d_m \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \phi_m(\mathbf{x}_i)^T \phi_m(\mathbf{x}_j) = \sum_{m=1}^{M} d_m \mathbf{c}_m^T \mathbf{c}_m \end{aligned}. \tag{9}$$

After substituting (8) with (9) into $\|\phi(\mathbf{x}) - \mathbf{c}\|^2 = k(\mathbf{x}, \mathbf{x}) - 2\mathbf{c}^T \phi(\mathbf{x}) + \mathbf{c}^T \mathbf{c}$, we have

$$\begin{aligned} \|\phi(\mathbf{x}) - \mathbf{c}\|^2 &= \sum_{m=1}^{M} d_m k_m(\mathbf{x}, \mathbf{x}) - 2\sum_{m=1}^{M} d_m \mathbf{c}_m^T \phi_m(\mathbf{x}) + \sum_{m=1}^{M} d_m \mathbf{c}_m^T \mathbf{c}_m = \sum_{m=1}^{M} d_m \left(k_m(\mathbf{x}, \mathbf{x}) - 2\mathbf{c}_m^T \phi_m(\mathbf{x}) + \mathbf{c}_m^T \mathbf{c}_m \right) \\ &= \sum_{m=1}^{M} d_m \left(\phi_m(\mathbf{x})^T \phi_m(\mathbf{x}) - 2\mathbf{c}_m^T \phi_m(\mathbf{x}) + \mathbf{c}_m^T \mathbf{c}_m \right) = \sum_{m=1}^{M} d_m \|\phi_m(\mathbf{x}) - \mathbf{c}_m\|^2 \end{aligned}. \tag{10}$$

Clearly, equation (10) shows the fact that the square distance between the data sample $\mathbf{x}$ and the sphere center $\mathbf{c}$, with a single kernel function k, can be expressed as the linear combination of the M square distances, each of which is computed by the data sample $\mathbf{x}$, the sphere center $\mathbf{c}_m$ with the base kernel k_m. This implies that, with multiple kernels, M spheres with centers $\mathbf{c}_m$ and optimized radius r can be obtained by the SSLM model in (4). Thus, we can define the margin ρ_m^2 $(1 \le m \le M)$ between the outliers and the surface of each sphere, and the total margin can be computed by the linear combination of the margins ρ_m^2; i.e.: $\rho^2 = \sum_{m=1}^{M} d_m \rho_m^2$.

3.2 MKSLM Dual Problem

The Lagrangian function of MKSLM in (7) can be formulated as follows:

$$L(r, \mathbf{c}_m, \xi_i, \boldsymbol{\alpha}, \boldsymbol{\beta}) = r^2 - \mu\sum_{m=1}^{M} d_m\rho_m^2 + \frac{1}{v_1 m_1}\sum_{i=1}^{m_1}\xi_i + \frac{1}{v_2 m_2}\sum_{j=m_1+1}^{N}\xi_j + \sum_{i=1}^{m_1}\alpha_i\left(\sum_{m=1}^{M} d_m\left\|\phi_m(\mathbf{x}_i) - \mathbf{c}_m\right\|^2 - r^2 - \xi_i\right) - \sum_{j=m_1+1}^{N}\alpha_j\left(\sum_{m=1}^{M} d_m\left\|\phi_m(\mathbf{x}_j) - \mathbf{c}_m\right\|^2 - r^2 - \sum_{m=1}^{M} d_m\rho_m^2 + \xi_j\right) - \sum_{i=1}^{N}\beta_i\xi_i \tag{11}$$

where $\boldsymbol{\alpha} = [\alpha_1,...,\alpha_N]^T$ and $\boldsymbol{\beta} = [\beta_1,\cdots,\beta_N]^T$ are the vector of Lagrange multipliers. Then, by setting the partial derivatives of $L(r,\mathbf{c}_m,\xi_i,\boldsymbol{\alpha},\boldsymbol{\beta})$ with respect to the primal variables to zero, the following formulae are obtained:

$$\frac{\partial L}{\partial r} = 2r\left(1 - \sum_{i=1}^{N}\alpha_i y_i\right) = 0, \tag{12}$$

$$\frac{\partial L}{\partial \mathbf{c}_m} = -2d_m\left(\sum_{i=1}^{N}\alpha_i y_i\phi_m(\mathbf{x}_i) + \sum_{i=1}^{N}\alpha_i y_i\mathbf{c}_m\right) = 0, \tag{13}$$

$$\frac{\partial L}{\partial \rho_m} = -2d_m\rho_m\left(\mu - \sum_{j=m_1+1}^{N}\alpha_j\right) = 0, \tag{14}$$

$$\frac{\partial L}{\partial \xi_i} = \frac{1}{v_1 m_1} - \alpha_i - \beta_i = 0, 1 \le i \le m_1, \tag{15}$$

$$\frac{\partial L}{\partial \xi_j} = \frac{1}{v_2 m_2} - \alpha_j - \beta_j = 0, m_1 + 1 \le j \le N. \tag{16}$$

Following simple algebraic steps, we have

$$\sum_{i=1}^{N}\alpha_i y_i = 1, \tag{17}$$

$$\mathbf{c}_m = \sum_{i=1}^{N}\alpha_i y_i\phi_m(\mathbf{x}_i), \tag{18}$$

$$\sum_{j=m_1+1}^{N}\alpha_j = \mu, \tag{19}$$

$$0 \le \alpha_i \le \frac{1}{v_1 m_1}, 1 \le i \le m_1, \tag{20}$$

$$0 \le \alpha_j \le \frac{1}{v_2 m_2}, m_1 + 1 \le j \le N. \tag{21}$$

By substituting (17), (18) and (19) to the Lagrangian function of MKSLM in (7), we can obtain its dual problem:

$$\begin{aligned} &\max_{\boldsymbol{\alpha}} \ (\boldsymbol{\alpha} \circ \mathbf{y})^T diag\left(\sum_{m=1}^{M} d_m \mathbf{K}_m\right) - (\boldsymbol{\alpha} \circ \mathbf{y})^T \left(\sum_{m=1}^{M} d_m \mathbf{K}_m\right)(\boldsymbol{\alpha} \circ \mathbf{y}) \\ &\text{s.t. } \mathbf{1}^T(\boldsymbol{\alpha} \circ \mathbf{y}) = 1, \ \frac{0.5}{v_2 m_2}(\mathbf{y} - \mathbf{1}) \le (\boldsymbol{\alpha} \circ \mathbf{y}) \le \frac{0.5}{v_1 m_1}(\mathbf{y} + \mathbf{1}), \ \mathbf{y}^T(\boldsymbol{\alpha} \circ \mathbf{y}) = 2\mu + 1. \end{aligned} \tag{22}$$

Given the kernel combination coefficient $\mathbf{d}$, the above QP problem can be solved. Then, an unknown data sample $\mathbf{x}$ can be detected by using the decision function below:

$$f(\mathbf{x}) = r^2 - \sum_{m=1}^{M} d_m \left\| \phi_m(\mathbf{x}) - \mathbf{c}_m \right\|^2, \tag{23}$$

where $r^2 = \frac{1}{|S_2|} \sum_{\mathbf{x}_i \in S_2} \sum_{m=1}^{M} d_m \left\| \phi_m(\mathbf{x}_i) - \mathbf{c}_m \right\|^2$, $S_2 = \{\mathbf{x}_i \mid 0 < \alpha_i < \frac{1}{v_1 m_1}, 1 \le i \le m_1\}$. $\mathbf{x}$ is a normal data sample if $f(\mathbf{x}) \ge 0$. Otherwise, it is an outlier, or an abnormal data sample. However, if the kernel combination coefficient $\mathbf{d}$ is unknown, the optimization problems in (7) and (22) cannot be solved. A possible method is to use the alternate optimization algorithms developed in2728. In the following section, another method is presented which is similar to the SimpleMKL algorithm 19 to solve the optimization problem.

3.3 Algorithm for Solving MKSLM

To obtain the optimized coefficients of base kernels, we consider the following constrained optimization problem:

$$\begin{aligned} &\min_{\mathbf{d}} \ J^{primal}(\mathbf{d}), \\ &\text{s.t. } \mathbf{1}^T \mathbf{d} = 1, \mathbf{d} \ge 0. \end{aligned} \tag{24}$$

where $J^{primal}(\mathbf{d})$ is the original problem of MKSLM associated with the optimized coefficients $\mathbf{d}$; i.e.:

$$J^{primal}(\mathbf{d}) = \begin{cases} \min\limits_{r,\mathbf{c}} \; r^2 - \mu\sum\limits_{m=1}^{M} d_m \rho_m^2 + \dfrac{1}{v_1 m_1}\sum\limits_{i=1}^{m_1}\xi_i + \dfrac{1}{v_2 m_2}\sum\limits_{j=m_1+1}^{N}\xi_j, \\ \text{s.t.} \; \sum\limits_{m=1}^{M} d_m \left\|\phi_m(\mathbf{x}_i) - \mathbf{c}_m\right\|^2 \le r^2 + \xi_i, 1 \le i \le m_1, \\ \quad \sum\limits_{m=1}^{M} d_m \left\|\phi_m(\mathbf{x}_j) - \mathbf{c}_m\right\|^2 \ge r^2 + \sum\limits_{m=1}^{M} d_m \rho_m^2 - \xi_j, \\ \quad m_1 + 1 \le j \le N, \; \xi_i \ge 0, 1 \le i \le N. \end{cases} \tag{25}$$

By using the Lagrangian technique, we can obtain the dual problem associated with the optimization problem in (25); i.e.:

$$J^{dual}(\mathbf{d}) = \begin{cases} \max\limits_{\boldsymbol{\alpha}} \; (\boldsymbol{\alpha}\circ\mathbf{y})^T diag\left(\sum\limits_{m=1}^{M} d_m \mathbf{K}_m\right) - (\boldsymbol{\alpha}\circ\mathbf{y})^T\left(\sum\limits_{m=1}^{M} d_m \mathbf{K}_m\right)(\boldsymbol{\alpha}\circ\mathbf{y}) \\ \text{s.t.} \; \mathbf{1}^T(\boldsymbol{\alpha}\circ\mathbf{y}) = 1, \; \dfrac{0.5}{v_2 m_2}(\mathbf{y}-\mathbf{1}) \le (\boldsymbol{\alpha}\circ\mathbf{y}) \le \dfrac{0.5}{v_1 m_1}(\mathbf{y}+\mathbf{1}), \mathbf{y}^T(\boldsymbol{\alpha}\circ\mathbf{y}) = 2\mu + 1. \end{cases} \tag{26}$$

Let $J(\mathbf{d})$ be the optimal objective value of the original problem in (25). Due to strong duality, $J(\mathbf{d})$ is also the objective value of the dual problem in (26); i.e.:

$$J(\mathbf{d}) = (\boldsymbol{\alpha}^*\circ\mathbf{y})^T diag\left(\sum_{m=1}^{M} d_m \mathbf{K}_m\right) - (\boldsymbol{\alpha}^*\circ\mathbf{y})^T\left(\sum_{m=1}^{M} d_m \mathbf{K}_m\right)(\boldsymbol{\alpha}^*\circ\mathbf{y}) \tag{27}$$

where $\boldsymbol{\alpha}^*$ maximizes (26). Then, we can compute the differentiation of $J(\mathbf{d})$ as follows:

$$\frac{\partial J(\mathbf{d})}{\partial d_m} = (\boldsymbol{\alpha}^*\circ\mathbf{y})^T diag(\mathbf{K}_m) - (\boldsymbol{\alpha}^*\circ\mathbf{y})^T(\mathbf{K}_m)(\boldsymbol{\alpha}^*\circ\mathbf{y}). \tag{28}$$

Using (28), the gradient direction of $J(\mathbf{d})$ can be obtained. This approach is similar to the gradient descent method19. To satisfy the equality and the non-negativity constraints in (4), i.e. $\mathbf{1}^T\mathbf{d} = 1$ and $\mathbf{d} \ge 0$, we compute the gradient direction of $J(\mathbf{d})$ using the Reduced Gradient

Algorithm19. Once the descent direction $\mathbf{g} = [g_m] \in \Re^M$ is computed, the kernel combination coefficient $\mathbf{d}$ can be updated iteratively as follows:

$$\mathbf{d}^{(t+1)} = \mathbf{d}^{(t)} + \eta^{(t)}\mathbf{g}^{(t)}, \tag{29}$$

where $\eta^{(t)}$ is the *t*th learning rate, $\mathbf{g}^{(t)}$ is the *t*th updating direction.

The algorithm for solving the proposed MKSLM is summarized in Algorithm1.

Algorithm 1. MKSLM
1. Set $\mathrm{d} = 1/M$;
2. Repeat
Solve the dual problem $J^{dual}(\mathbf{d})$ of MKSLM using (26);
Compute the gradient $\nabla J(\mathbf{d})$ of $J(\mathbf{d})$ using (28);
Compute the descent direction $\mathbf{g}^{(t)}$ using the Reduced Gradient Algorithm;
Update the kernel combination coefficient $\mathbf{d}$ using (29);
Until stopping criterion is met.
3. End.

4. Experiment Results

The performance of the proposed MKSLM is evaluated by comparing it with that of a number of related methods, including SVM 12, SVM-C29, SVDD 1 and SSLM11. SVM is a popular classification algorithm while SVM-C is a modified SVM model used to deal with the problem with imbalanced data sets 29. SVDD and SSLM are two widely used novelty detection approaches that have been discussed in section 2. Twelve data sets are employed to conduct the performance evaluation experiments.

4.1 Data Preparation

The nine data sets adopted in the experiments are obtained from the Pattern Recognition Laboratory of the Delft University of Technology, the Netherlands. The details of data sets are summarized in Table 1. To prepare a training data set, m_1 samples were randomly chosen from 70 percent of the targets, and a small number of outliers were introduced, such that 95 percent of the

training data are the targets and the remaining 5 percent are outliers; i.e., $m_1/(m_1+m_2)\approx 0.95$. Before the experiments, each data set was normalized with respect to each feature to yield unit norm.

Table 1 Property of Data Sets in Experiment

Data sets	The number of targets	The number of outliers	Dimension	m_1	m_2
Biomed	127	67	5	89	5
Iris	50	100	4	35	2
Spectf	95	254	44	67	4
Liver	200	145	6	0	7
Wine	59	119	13	41	2
Thyriod	93	3679	21	65	3
Breast	458	241	9	1	17
Ionosphere	225	126	34	8	8
Sonar	111	97	60	78	4

4.2 Parameters Setting

For MKSLM, we used three kernels, the Gaussian kernel $k(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|^2/\gamma_G)$, the Laplacian kernel $k(\mathbf{x}_i,\mathbf{x}_j)=\exp(-\|\mathbf{x}_i-\mathbf{x}_j\|/\gamma_L)$, and the PolyPlus kernel $k(\mathbf{x}_i,\mathbf{x}_j)=(1+\mathbf{x}_i^T\mathbf{x}_j)^d$. Nine kernel parameters $\gamma_G=2^\tau\sigma^2$ were used for the Gaussian kernels and another nine kernel parameters $\gamma_L=2^\tau\sigma$ were used for the Laplacian kernels, where $\tau\in\{-4,-3,-2,-1,0,1,2,3,4\}$ and σ is the mean L2-norm of the training data. For PolyPlus kernel, three kernel parameters; i.e., $d\in\{1,2,3\}$, were used. Therefore, a total of 21 base kernels were adopted in the MKSLM algorithm. For the other algorithms employed for comparison, the Gaussian kernel was adopted and the best value of its parameter was searched from the grid $\sigma^2\times\{1/16,1/8,1/4,1/2,1,2,4,8,16\}$.

For MKSLM and SSLM, the best value for the parameter μ was searched from the grid {0.1,0.5,1,5,10,30,50,70,90}, while the parameters v_1 and v_2 were selected from {0.01,0.001}. For SVM, the best value for the parameter C was searched from the grid {0.01,0.05,0.1,0.5,1,5,10,50,100,500}. For SVM-C and SVDD, the parameter C_1 was searched from the same grid as SVM, and C_2 was selected from the values with respect to the grid $C_2/C_1 \in \{\frac{1}{4}\times\frac{m_1}{m_2},\frac{1}{2}\times\frac{m_1}{m_2},\frac{m_1}{m_2},2\times\frac{m_1}{m_2},4\times\frac{m_1}{m_2}\}$.

4.3 Detection Results

Considering the imbalance of the data sets, the geometric mean metric was employed in order to measure the performance of the compared algorithms. This metric takes into consideration the classification results of both positive and negative classes, and it is computed as: $g=\sqrt{a^+\cdot a^-}$, where a^+ and a^- are defined as follows:

$$a^+=\frac{\#\text{ targets correctly classified}}{\#\text{total targets classified}}\times 100\%,\qquad a^-=\frac{\#\text{ outliers correctly classified}}{\#\text{total outliers classified}}\times 100\%.$$

All algorithms were implemented using Matlab. We performed cross-validation on the training data for selecting the parameters. The experiment on each data set was repeated ten times, and the mean and standard deviation of the geometric mean metric were obtained. The results are shown in Table 2, where the values in bold face indicate the method showing the best performance. It can be seen that MKSLM and SSLM achieve the best performance. Furthermore, the performance of SVM was the worst, which validates that it is unsuitable for classifying imbalanced data sets. Although SVM-C modifies the SVM model in an attempt to solve the learning problem of imbalanced data sets, it is still inferior to SVDD, SSLM and MKSLM.

Table 2 Outliers Detection Results of 12 Data Sets

Data sets	SVM	SVM-C	SVDD	SSLM	MKSLM
Biomed	73.17±6.13	79.34±3.80	65.40±1.51	**82.24±2.77**	78.87±5.69
Iris	89.09±8.52	84.43±9.99	86.16±9.49	93.98±5.94	**94.25±2.87**
Spectf	44.24±23.61	56.01±12.91	**77.11±3.16**	73.54±6.07	74.89±4.45
Liver	36.57±5.68	37.50±7.99	**62.37±6.14**	58.42±5.60	60.45±3.09
Wine	81.36±8.51	81.41±7.01	95.71±4.63	93.42±2.96	**96.56±3.23**

Thyriod	68.84±25.03	85.55±4.27	80.01±5.90	87.28±4.30	**91.51±3.58**
Breast	83.95±4.56	81.24±5.48	95.05±1.44	**95.27±1.10**	93.80±2.03
Ionosphere	62.91±8.76	65.28±5.07	88.82±2.71	**91.79±2.09**	90.07±3.63
Sonar	35.44±11.92	50.29±12.14	65.51±2.01	69.01±4.01	**70.23±4.07**

5. Discussion

Since SSLM explicitly presents the classification margin in geometry, the multiple kernel technique can be integrated with it and the presented MKSLM in this paper can better capture the surface of the sphere and detect the outliers. However, the original SVDD does not geometrically provide the classification margin. Then, the multiple kernel technique cannot be used in SVDD. While SVM or v-SVM inherently obtain the classification margin when it finds the optimal hyperplane, the multiple kernel technique can be introduced easily, e.g. SimpleMKL19. Thus, the classification margin plays important roles in the multiple kernel technique.

6. Conclusion

This paper proposes the new method MKSLM for novelty detection. The method uses multiple base kernels to maximize the margin between the surface of the sphere and the outliers during constructing a small sphere. In comparison with SVM, SVM-C, SVDD and SSLM, the experimental results show that this method can handle imbalanced data and achieve significant performance. 19In this paper, the multiple kernel technique is currently being extended to other machine learning methods.

Acknowledgement

This work was supported in part by Zhejiang Provincial Natural Science Foundation of China under Grant LY14F010010; the National Natural Science Foundation of China under Grants 60872057, 61170122, 61170029, 61272210, 61202311.

References

1. Tax D M J, Duin R P W,Support vector data description, 2004, Machine Learning, vol.54, no. 1, pp.45-66.
2. Schölkopf B, Smola A J, Learning with Kernels, 2002, Cambridge, MA: MIT Press.

3. Campbell C, Bennett K P, A linear programming approach to novelty detection, 2000, Advances in Neural Information Processing Systems 13, Cambridge, MA: MIT Press, pp. 395-401.
4. Yin S, Zhu X P, Jing C. Fault detection based on a robust one class support vector machine, 2014, Neurocomputing, vol.145, pp.263-268.
5. Yin G, Zhang Y T, Li Z N, Ren G Q, Fan H B, Online fault diagnosis method based on Incremental Support Vector Data Description and Extreme Learning Machine with incremental output structure, 2014, Neurocomputing, vol.128, pp.224-231.
6. Hajji H, Statistical analysis of network traffic for adaptive faults detection, 2005, IEEE Transactions on Neural Networks, vol.16, no.5, pp.1053-1063.
7. Chen Y X, Dang X, Peng H X, Bart H L, Outlier detection with the kernelized spatial depth function, 2009, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, no.2, pp.288-305.
8. Moya M, Hush D, Network constraints and multi-objective optimization for one-class classification, 1996, Neural Networks, vol.9, no.3, pp.463-474.
9. Towel G G, Local expert autoassociators for anomaly detection, 2000, In Proceedings of the Seventeenth International Conference on Machine Learning, San Francisco, CA: Morgan Kaufmann Publishers, pp.1023-1030.
10. Roberts, S, Tarassenko L, A probabilistic resource allocation network for novelty detection, 1994, Neural Computation, vol.6, no.2, pp. 270-284.
11. Wu M R, Ye J P, A small sphere and large margin approach for novelty detection using training data with outliers, 2009, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.31, no.11, pp. 2088-2092.
12. Cortes C, Vapnik V N, Support vector networks, 1995, Machine Learning, vol.20, no.3, pp. 273-297.
13. Schölkopf B, Smola A, Williamson R C, Bartlett P L, New support vector algorithms, 2000, Neural Computation, vol.12, no.5, pp.1207-1245.
14. Cha M, Kim J S, Baek J G, Density weighted support vector data description, 2014, Expert Systems with Applications, vol.41, no.7, pp. 3343-3350.
15. Wang C D, Lai J J, Position regularized Support Vector Domain Description, 2013, Pattern Recognition, vol.46, no.3, pp. 875-884.

16. Lanckriet G, Cristianini N, Bartlett P, Ghaoui L E, Jordan M I, Learning the kernel matrix with semidefinite programming, 2004, Journal of Machine Learning Research, vol.5, pp. 27-72.
17. Bach F R, Lanckriet G R G, Jordan M, Multiple kernel Learning, conic duality, and the SMO algorithm, 2004, In Proceedings of the Twenty-first International Conference on Machine Learning, New York, NY: ACM.
18. Sonnenburg S, Rätsch G, Schäfer C, Schölkopf B, Large scale multiple kernel learning, 2006, Journal of Machine Learning Research, vol.7, pp.1531-1565.
19. Rakotomamonjy A, Bach F R, Canu S, Grandvalet Y, SimpleMKL, 2008, Journal of machine Learning Research, vol.9, pp. 2491-2521.
20. Hu M, Chen Y, Kwok J T, Building sparse multi-kernel SVM classifiers, 2009, IEEE Transactions on Neural Networks, vol.20, no.5, pp. 827-839.
21. Lin Y Y, Liu T L, Fuh C S, Multiple kernel learning for dimensionality reduction, 2011, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.33, no.6, pp. 1147-1160.
22. Duan L X, Tsang I W, Xu D, Domain transfer multiple kernel learning, 2012, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.34, no.3, pp. 465-479.
23. Sun J, Wu X, Yan S, Cheong L F, Chua T S, Li J, Hierarchical spatio-temporal context modeling for action recognition, 2009, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), pp. 2004-2011.
24. Wu X, Xu D, Duan L, Luo J, Action recognition using context and appearance distribution features, 2011, In Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2011), pp. 489-496.
25. Vedaldi A, Gulshan V, Varma M, Zisserman A, Multiple kernels for object detection, 2009, In Proceedings of IEEE Conference on Computer Vision (ICCV 2009), pp.606-613.
26. Wang J J Y, Huang J H Z, Sun Y J, Gao X, Feature selection and multi-kernel learning for adaptive graph regularized nonnegative matrix factorization, 2015, Expert Systems with Applications, vol.42, no.3, pp.1278-1286.
27. Grandvalet Y, Canu S, Adaptive scaling for feature selection in SVMs, 2002, Advances in Neural Information Processing Systems 15 (NIPS 2002), Cambridge, MA: MIT Press, pp.553-560.
28. Grandvalet Y, Canu S, Outcomes of the equivalence of adaptive ridge with least absolute shrinkage, 1998, Advances in Neural Information Processing Systems 11, pp.445-451.

29. Veropoulos K, Campbell C, Cristianini N, Controlling the sensitivity of support vector machines, 1999, Proceeding in International Joint Conference on Artificial Intelligence, pp.55-60.