Fuzzy Multi-Sphere Support Vector Data Description

Trung Le[1], Dat Tran[2], and Wanli Ma[2]

[1] HCM City University of Education, Vietnam
[2] University of Canberra, ACT 2601, Australia
dat.tran@canberra.edu.au

Abstract. Current well-known data description methods such as Support Vector Data Description and Small Sphere Large Margin are conducted with assumption that data samples of a class in feature space are drawn from a single distribution. Based on this assumption, a single hypersphere is constructed to provide a good data description for the data. However, real-world data samples may be drawn from some distinctive distributions and hence it does not guarantee that a single hypersphere can offer the best data description. In this paper, we introduce a Fuzzy Multi-sphere Support Vector Data Description approach to address this issue. We propose to use a set of hyperspheres to provide a better data description for a given data set. Calculations for determining optimal hyperspheres and experimental results for applying this proposed approach to classification problems are presented.

Keywords: Kernel Methods, Fuzzy Interference, Support Vector Data Description, Multi-Sphere Support Vector Data Description.

1 Introduction

Support Vector Machine (SVM) [2], [4] has been proven a very effective method for binary classification. However, it cannot render good performance for one-class classification problems where one of two classes is under-sampled, or only data samples of one class are available for training [9] . One-class classification involves learning data description of normal class to build a model that can detect any divergence from normality [11]. The samples of abnormal class if existed contribute to refining the data description. Support Vector Data Description (SVDD) was introduced in [14], [13] as a kernel method for one-class classification. SVDD aims at constructing an optimal hypersphere in feature space which includes only normal samples and excludes all abnormal samples with tolerances. This optimal hypersphere is regarded as a data description since when mapped back to input space it becomes a set of contours that tightly enclose the normal data samples [1].

Variations of SVDD were proposed to enhance this approach. In [9], density-induced information was incorporated to the samples so that the dependency of data description on support vectors can be less imposed when these support

J. Pei et al. (Eds.): PAKDD 2013, Part II, LNAI 7819, pp. 570–581, 2013.

vectors cannot characterise well the data. To reduce the impact of less important dimensions, a single ellipse was learnt rather than a single hypersphere [10]. However, this work was not general since it was only proposed for the model in input space. Other approaches introduced better margins for SVDD such as [15] [7]. To reduce the chance of acceptance of outliers, in [15] a small sphere with large margin was proposed. However, this can induce side-effect which causes the interference of decision boundary into normal data region. To overcome this issue, an optimal sphere with two adjustable margins for reducing both true positive ($TP\%$) and true negative ($TN\%$) error rates was proposed [7].

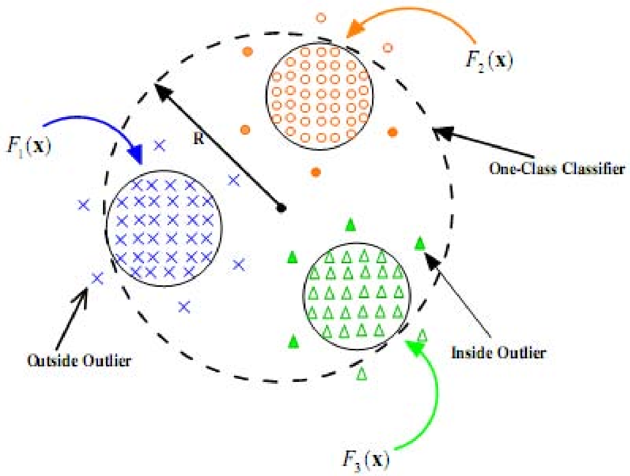

Fig. 1. Inside outliers would be improperly included if only one hypersphere is constructed [16]

SVDD assumes that all samples of the training set are drawn from a single uniform distribution [13]. However, this hypothesis is not always true since real-world data samples may be drawn from distinctive distributions [5]. Therefore, a single hypersphere cannot be a good data description. For example, in Figure 1, data samples are scattered over some distinctive distributions and one single hypersphere would improperly record the inside outliers. In [16], a multi-sphere approach to SVDD was proposed for multi-distribution data. The domain for each distribution was detected and for each domain an optimal sphere was constructed to describe the corresponding distribution. However, the learning process was heuristic and did not follow up learning with minimal volume principle [12]. In [6], a method was proposed to link the input space to the feature space. The dense regions (clusters) in the input space were identified and became a single sphere in the feature space. Again, this method was heuristic and did not abide by learning with minimum volume principle. To motivate learning with this principle, a hard multi-sphere support vector data description (HMS-SVDD) [8] was proposed. A set of hyperspheres was introduced to enclose all the data samples. A data sample will belong to only one hypersphere. This restriction should be relaxed to allow a data sample to belong to different hyperspheres if that sample has similar degrees of belonging to those hyperspheres. To address this issue, we propose fuzzy multi-sphere support vector data description (FMS-SVDD) in this paper. A fuzzy membership is assigned to each data sample to denote the degree of belonging of that sample to a hypersphere. We prove that classification error will be reduced after each iteration in the learning process. The set of hyperspheres will gradually converge to a stable configuration. To evaluate the proposed apprach, we performed classification experiments on 23 data sets in UCI repository. The experimental results showed that FMS-SVDD could provide better classification rates in comparison to other one-class kernel methods.

2 Fuzzy Multi-Sphere Support Vector Data Description (FMS-SVDD)

2.1 Problem Formulation

Let $x_i,\ i = 1,\ldots,p$ be normal data samples with label $y_i = +1$ and $x_i,\ i = p+1,\ldots,n$ be abnormal data samples with label $y_i = -1$. Consider a set of m hyperspheres $S_j(c_j, R_j)$ with center c_j and radius $R_j,\ j = 1,\ldots,m$. This hypershere set is a good data description of the data set $X = \{x_1, x_2, \ldots, x_n\}$ if each of the hyperspheres describes a distribution in this data set and the sum of all radii $\sum_{j=1}^{m} R_j^2$ should be minimised.

Let matrix $U = [u_{ij}]_{p\times m},\ u_{ij} \in [0,1],\ i = 1,\ldots,p,\ j = 1,\ldots,m$ where u_{ij} is the membership representing degree of belonging of sample x_i to hypersphere S_j. It is necessary to construct a set of hyperspheres so that these hyperspheres can include all normal data and exclude all abnormal data with tolerances. The optimisation problem of fuzzy multi-sphere SVDD can be formulated as follows

$$\min_{R,c,\xi,u}\left(\sum_{j=1}^{m} R_j^2 + \frac{1}{\nu_1 p}\sum_{i=1}^{p}\xi_i + \frac{1}{\nu_2 q}\sum_{i=p+1}^{n}\sum_{j=1}^{m}\xi_{ij}\right) \tag{1}$$

subject to

$$\begin{aligned} &\sum_{j=1}^{m} u_{ij}^d \|\phi(x_i) - c_j\|^2 \le \sum_{j=1}^{m} u_{ij}^d R_j^2 + \xi_i,\ i = 1,\ldots,p \\ &\|\phi(x_i) - c_j\|^2 \ge R_j^2 - \xi_{ij},\ i = p+1,\ldots,n,\ j = 1,\ldots,m \\ &\sum_{j=1}^{m} u_{ij} = 1,\ i = 1,\ldots,p \end{aligned} \tag{2}$$

where $R = [R_j]_{j=1,\ldots,m}$ is vector of radii, ν_1 and ν_2 are constants, ξ_i and ξ_{ij} are slack variables, $c = [c_j]_{j=1,\ldots,m}$ is vector of centres, and $q = n - p$ is the number of abnormal (negative) samples.

In the FMS-SVDD model, we also introduce parameter $d > 1$ to adjust the relative ratios among the memberships u_{ij}, $i = 1, \ldots, p$, $j = 1, \ldots, m$.

Minimising the function in (1) over variables R, c and ξ subject to (2) will determine radii and centres of hyperspheres and slack variables if the matrix U is given. On the other hand, the matrix U will be determined if radii and centres of hyperspheres are given. Therefore an iterative algorithm will be applied to find a complete solution. The algorithm consists of two alternative steps: *1) Calculate radii and centres of hyperspheres and slack variables*, and *2) Calculate membership U.*

We present in the next sections the iterative algorithm and the proof of key theorem which states that the classification error in the current iteration will be smaller than that in the previous iteration. It means that the model is gradually refined and converged to a stable configuration.

For classifying a sample x, the following decision function is used:

$$f(x) = sign\Big(\max_{1 \leq j \leq m} \left\{ R_j^2 - ||\phi(x) - c_j||^2 \right\} \Big) \tag{3}$$

The unknown sample x is normal if $f(x) = +1$ or abnormal if $f(x) = -1$. This decision function implies that the mapping of a normal sample has to be in one of the hyperspheres and that the mapping of an abnormal sample has to be outside all of those hyperspheres.

The following theorem is used to consider the relation of slack variables to the classified samples:

Theorem 1. *Assume that (R, c, ξ_i, ξ_{ij}) is a solution of the optimisation problem (1), x_i, $i \in \{1, 2, \ldots, n\}$ is the i-th sample. The slack variable ξ_i or ξ_{ij} can be computed as*

$$\begin{aligned} \xi_i &= \max \left\{ 0, \sum_{j=1}^{m} u_{ij}^d \left(\|\phi(x_i) - c_j\|^2 - R_j^2 \right) \right\},\ i = 1, \ldots, p \\ \xi_{ij} &= \max \left\{ 0, R_j^2 - \|\phi(x_i) - c_j\|^2 \right\},\ i = p+1, \ldots, n,\ j = 1, \ldots, m \end{aligned} \tag{4}$$

Proof

For all i, from equation (2) we have

$$\begin{aligned} \xi_i &\geq \max \left\{ 0, \sum_{j=1}^{m} u_{ij}^d \left(\|\phi(x_i) - c_j\|^2 - R_j^2 \right) \right\},\ i = 1, \ldots, p \\ \xi_{ij} &\geq \max \left\{ 0, R_j^2 - \|\phi(x_i) - c_j\|^2 \right\}, \quad i = p+1, \ldots, n,\ j = 1, \ldots, m \end{aligned} \tag{5}$$

Moreover, (R, c, ξ) is minimal solution of (1). Hence, the *theorem 1* is proved.

It is natural to define $error(i)$, i.e. the error at sample x_i, $1 \leq i \leq n$ as follows

If *x_i is normal data sample* **then**

$$error(i) = \begin{cases} 0 & if\, x_i\, is\, correctly\, classified \\ \min\limits_{1 \leq j \leq m} \left\{ \|\phi(x_i) - c_j\|^2 - R_j^2 \right\} & otherwise \end{cases} \tag{6}$$

Else

$$error(i) = \begin{cases} 0 & if\, x_i\, is\, correctly\, classified \\ \min\limits_{j \in J} \left\{ R_j^2 - \|\phi(x_i) - c_j\|^2 \right\} & otherwise \end{cases} \tag{7}$$

where $J = \{j : x_i \in S_j \text{ and } 1 \leq j \leq m\}$.

We can prove that $\sum\limits_{i=1}^{p} \xi_i$ is an upper bound of $\frac{1}{m^{d-1}} \sum\limits_{i=1}^{p} error(i)$, and $\sum\limits_{i=p+1}^{n} \sum\limits_{j=1}^{m} \xi_{ij}$ is an upper bound of $\sum\limits_{i=p+1}^{n} error(i)$. The second inequality is trivial, therefore we primarily concentrate on the first one.

Theorem 2. *$\sum\limits_{i=1}^{p} \xi_i$ is an upper bound of $\frac{1}{m^{d-1}} \sum\limits_{i=1}^{p} error(i)$.*

Proof

Let us denote $d_{ij} = \|\phi(x_i) - c_j\|^2 - R_j^2,\ i = 1, \ldots, p,\ j = 1, \ldots, m$. Given $1 \leq i \leq p$, we will prove that $\xi_i \geq \frac{1}{m^{d-1}} error(i)$. This above inequality is trivial if x_i is correctly classified. We consider the case where x_i is misclassified, i.e., $d_{ij} > 0$ for all j. It means that $\xi_i = \sum\limits_{j=1}^{m} u_{ij}^d d_{ij}$. From definition of $error(i)$, we have: $\xi_i = \sum\limits_{j=1}^{m} u_{ij}^d d_{ij} \geq error(i) \sum\limits_{j=1}^{m} u_{ij}^d$. To fulfill this proof, we will show that $\sum\limits_{j=1}^{m} u_{ij}^d \geq \frac{1}{m^{d-1}} \left(\sum\limits_{j=1}^{m} u_{ij} \right)^d = \frac{1}{m^{d-1}}$. Indeed, this inequality can be rewritten as follows

$$\sum_{j=1}^{m} \left(\frac{m u_{ij}}{\sum\limits_{j'=1}^{m} u_{ij'}} \right)^d \geq m \tag{8}$$

By referring to Bernoulli inequality which says $(1+x)^r \geq 1 + rx$ if $r \geq 1$ and $x > -1$, we have

$$\left(\frac{m u_{ij}}{\sum\limits_{j'=1}^{m} u_{ij'}} \right)^d = \left(1 - \left(1 - \frac{m u_{ij}}{\sum\limits_{j'=1}^{m} u_{ij'}} \right) \right)^d \geq 1 - d \left(1 - \frac{m u_{ij}}{\sum\limits_{j'=1}^{m} u_{ij'}} \right) \ for\ all\ j \tag{9}$$

It follows that

$$\sum_{j=1}^{m} \left(\frac{m u_{ij}}{\sum\limits_{j'=1}^{m} u_{ij'}} \right)^d \geq \sum_{j=1}^{m} \left(1 - d \left(1 - \frac{m u_{ij}}{\sum\limits_{j'=1}^{m} u_{ij'}} \right) \right) = m \tag{10}$$

2.2 Calculating Radii, Centres and Slack Variables

The Lagrange function for the optimisation problem (1) subject to (2) is as follows

$$\begin{aligned} L\left(R, c, \xi, \alpha, \beta\right) = & \sum_{j=1}^{m} R_j^2 + C_1 \sum_{i=1}^{p} \xi_i + C_2 \sum_{i=p+1}^{n} \sum_{j=1}^{m} \xi_{ij} \\ & + \sum_{i=1}^{p} \alpha_i \left(\sum_{j=1}^{m} u_{ij}^d \left(\|\phi(x_i) - c_j\|^2 - R_j^2 \right) - \xi_i \right) \\ & - \sum_{i=p+1}^{n} \sum_{j=1}^{m} \alpha_{ij} \left(\|\phi(x_i) - c_j\|^2 - R_j^2 + \xi_{ij} \right) \\ & - \sum_{i=p+1}^{n} \sum_{j=1}^{m} \beta_{ij}\xi_{ij} - \sum_{i=1}^{p} \beta_i \xi_i \end{aligned} \tag{11}$$

where $C_1 = \frac{1}{\nu_1 p}$, $C_2 = \frac{1}{\nu_2 q}$ and $q = n - p$, q is the number of abnormal data samples.

Setting derivatives of $L(R, c, \xi, \alpha, \beta)$ with respect to primal variables to 0, we obtain

$$\frac{\partial L}{\partial R_j} = 0 \rightarrow \sum_{i=1}^{p} u_{ij}^d \alpha_i - \sum_{i=p+1}^{n} \alpha_{ij} = 1,\ j = 1, \ldots, m \tag{12}$$

$$\frac{\partial L}{\partial c_j} = 0 \rightarrow c_j = \sum_{i=1}^{p} u_{ij}^d \alpha_i \phi(x_i) - \sum_{i=p+1}^{n} \alpha_{ij} \phi(x_i),\ j = 1, \ldots, m \tag{13}$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow \alpha_i + \beta_i = C_1,\ i = 1, \ldots, p \tag{14}$$

$$\frac{\partial L}{\partial \xi_{ij}} = 0 \rightarrow \alpha_{ij} + \beta_{ij} = C_2,\ i = p+1, \ldots, n \quad j = 1, \ldots, m \tag{15}$$

To get the dual form, we substitute equations (12)-(15) to the Lagrange function in (11) and obtain the following

$$\begin{aligned} & L(R, c, \xi, \alpha, \beta) = \sum_{i=1}^{p} \sum_{j=1}^{m} \alpha_i u_{ij}^d \left(\|\phi(x_i) - c_j\|^2 \right) - \sum_{i=p+1}^{n} \sum_{j=1}^{m} \alpha_{ij} \left(\|\phi(x_i) - c_j\|^2 \right) \\ & = \sum_{i=1}^{p} \sum_{j=1}^{m} u_{ij}^d \alpha_i K(x_i, x_i) - \sum_{j=1}^{m} \sum_{i=p+1}^{n} \alpha_{ij} K(x_i, x_i) \\ & -2 \sum_{j=1}^{m} c_j \left(\sum_{i=1}^{p} u_{ij}^d \alpha_i \phi(x_i) - \sum_{i=p+1}^{n} \alpha_{ij} \phi(x_i) \right) + \sum_{j=1}^{m} \|c_j\|^2 \left(\sum_{i=1}^{p} u_{ij}^d \alpha_i - \sum_{i=p+1}^{n} \alpha_{ij} \right) \\ & = \sum_{i=1}^{p} \alpha_i K(x_i, x_i) \sum_{j=1}^{m} u_{ij}^d - \sum_{j=1}^{m} \sum_{i=p+1}^{n} \alpha_{ij} K(x_i, x_i) - \sum_{j=1}^{m} \|c_j\|^2 \\ & = \sum_{i=1}^{p} \alpha_i s_i K(x_i, x_i) - \sum_{r,j} \alpha_{rj} K(x_r, x_r) - \sum_{j=1}^{m} \left\| \sum_{i=1}^{p} u_{ij}^d \alpha_i \phi(x_i) - \sum_{i=p+1}^{n} \alpha_{ij} \phi(x_i) \right\|^2 \\ & = \sum_{i=1}^{p} \alpha_i s_i K(x_i, x_i) - \sum_{r,j} \alpha_{rj} K(x_r, x_r) - \sum_{i,i'} u_i u_{i'} K(x_i, x_{i'}) \alpha_i \alpha_{i'} \\ & - \sum_{rj, r'j} \alpha_{rj} \alpha_{r'j} K(x_r, x_{r'}) + 2 \sum_{i,rj} u_{ij}^d K(x_i, x_r) \alpha_i \alpha_{rj} \end{aligned} \tag{16}$$

where $1 \leq i, i' \leq p$, $p+1 \leq r, r' \leq n$, $1 \leq j \leq m$, $u_i = [u_{i1}^d, u_{i2}^d, \ldots, u_{im}^d]$, $u_i u_{i'} = \sum_{j=1}^{m} u_{ij}^d u_{i'j}^d$, and $s_i = \sum_{j=1}^{m} u_{ij}^d$.

We come up with the following optimisation problem

$$\min_{\alpha} \left(\begin{array}{l} \sum_{i,i'} u_i u_{i'} K(x_i, x_{i'}) \alpha_i \alpha_{i'} + \sum_{rj,r'j} \alpha_{rj} \alpha_{r'j} K(x_r, x_{r'}) - 2 \sum_{i,rj} u_{ij}^d K(x_i, x_r) \alpha_i \alpha_{rj} \\ - \sum_{i=1}^{p} \alpha_i u_i K(x_i, x_i) + \sum_{r,j} \alpha_{rj} K(x_r, x_r) \end{array} \right) \tag{17}$$

subject to

$$\begin{array}{l} \sum_{i=1}^{p} u_{ij}^d \alpha_i - \sum_{i=p+1}^{n} \alpha_{ij} = 1, \ j = 1, \ldots, m \\ 0 \leq \alpha_i \leq C_1, \ i = 1, \ldots, p \\ 0 \leq \alpha_{ij} \leq C_2, \ i = p+1, \ldots, n, \ j = 1, \ldots, m \end{array} \tag{18}$$

Note that the number of variables in solution of the optimisation problem (17) is $p + (n-p) \times m$. In practice, we apply the Interior Point (IP) method [3] to solve out the above optimisation problem. The complexity is dependent on double logarithmic of tolerance ϵ, i.e. $log(log(1/\epsilon))$.

2.3 Calculating Membership *U*

We are in position to describe how to evaluate matrix U after obtaining new (R, c). Given $1 \leq i \leq p$, let us denote

$$\begin{array}{l} d_{ij} = \|\phi(x_i) - c_j\|^2 - R_j^2 \quad and \quad D_{ij} = \left(\frac{1}{d_{ij}}\right)^{\frac{1}{d-1}} \\ j_0 = \arg\min_{1 \leq j \leq m} d_{ij} \end{array} \tag{19}$$

The membership matrix can be updated as follows

If $d_{ij_0} \leq 0$ **then** $u_{ij_0} = 1$ **and** $u_{ij} = 0, \ j \neq j_0$
Else

$$u_{ij} = \frac{D_{ij}}{\sum_{k=1}^{m} D_{ik}}, \ j = 1, \ldots, m \tag{20}$$

2.4 Iterative Learning Process

The proposed iterative learning process for FMS-SVDD will run two alternative steps until a convergence is reached as follows

Initialise U by clustering the normal data set in the input space
Repeat *the following*
  Calculate R, c and ξ using U
  Calculate U using R and c
Until *convergence is reached*

2.5 Theoretical Background of FMS-SVDD

In the objective function $\sum_{j=1}^{m} R_j^2 + \frac{1}{\nu_1 p} \sum_{i=1}^{p} \xi_i + \frac{1}{\nu_2 q} \sum_{i=p+1}^{n} \sum_{j=1}^{m} \xi_{ij}$, the first summand can be regarded as regularisation quantity and the rest can be considered as empirical risk (referred to Theorem 2). We will prove that structural risk $\sum_{j=1}^{m} R_j^2 + \frac{1}{\nu_1 p} \sum_{i=1}^{p} \xi_i + \frac{1}{\nu_2 q} \sum_{i=p+1}^{n} \sum_{j=1}^{m} \xi_{ij}$ gradually becomes smaller in Theorem 4.

Theorem 3. *Given a m-multivariate function* $f(x_1, x_2, ..., x_m) = \sum_{i=1}^{m} d_i x_i^d$ *and* $d > 1$*. The following optimisation problem*

$$\min_x f(x) \tag{21}$$

subject to

$$\sum_{i=1}^{m} x_i = 1 \tag{22}$$

yields the solution as follows

If $d_{i_0} \le 0$ then $x_{i_0} = 1$ and $x_i = 0$, $i \ne i_0$
Else $x_i = \frac{\left(\frac{1}{d_i}\right)^{\frac{1}{d-1}}}{\sum_{k=1}^{m} \left(\frac{1}{d_k}\right)^{\frac{1}{d-1}}}, \quad i = 1, \ldots, m$
where $i_0 = \arg\min_{1 \le i \le m} d_i$.

Proof
Case 1**:** $d_{i_0} < 0$

$$\sum_{i=1}^{m} d_i x_i^d \ge d_{i_0} \sum_{i=1}^{m} x_i^d \ge d_{i_0} \sum_{i=1}^{m} x_i = d_{i_0} = f(0, ..., 1_{i_0}, ..., 0) \tag{23}$$

since $d > 1$.

Case 2**:** $d_{i_0} \ge 0$
The Lagrange function is of

$$L(x, \lambda) = \sum_{i=1}^{m} d_i x_i^d - \lambda \left(\sum_{i=1}^{m} x_i - 1 \right) \tag{24}$$

where λ is Lagrange multiplier.

Setting derivatives to 0, we gain

$$\begin{aligned} &\frac{\partial L}{\partial x_i} = 0 \Rightarrow d d_i x_i^{d-1} - \lambda = 0 \\ &\Rightarrow x_i = \left(\frac{\lambda}{d d_i}\right)^{\frac{1}{d-1}}, \ i = 1, \ldots, m \end{aligned} \tag{25}$$

From $\sum_{i=1}^{m} x_i = 1$, we have $x_i = \frac{\left(\frac{1}{d_i}\right)^{\frac{1}{d-1}}}{\sum_{k=1}^{m}\left(\frac{1}{d_k}\right)^{\frac{1}{d-1}}}, \quad i = 1, \ldots, m.$

Theorem 4. *Let ($R^{(t)}$, $c^{(t)}$, $\xi_i^{(t)}$, $\xi_{ij}^{(t)}$, $U^{(t)}$) and ($R^{(t+1)}$, $c^{(t+1)}$, $\xi_i^{(t+1)}$, $\xi_{ij}^{(t+1)}$, $U^{(t+1)}$) be solutions at the previous iteration and current iteration, respectively. The following inequality holds*

$$\begin{aligned}\sum_{j=1}^{m}\left(R_j^{(t+1)}\right)^2 + \frac{1}{\nu_1 p}\sum_{i=1}^{p}\xi_i^{(t+1)} + \frac{1}{\nu_2 q}\sum_{i=p+1}^{n}\sum_{j=1}^{m}\xi_i^{(t+1)} \leq \sum_{j=1}^{m}\left(R_j^{(t)}\right)^2 + \frac{1}{\nu_1 p}\sum_{i=1}^{p}\xi_i^{(t)} \\ + \frac{1}{\nu_2 q}\sum_{i=p+1}^{n}\sum_{j=1}^{m}\xi_i^{(t)}\end{aligned} \tag{26}$$

Proof
By referring to Theorem 3, it is easy to see that $u_i^{(t+1)} = (u_{i1}^{(t+1)}, ..., u_{im}^{(t+1)})$, $i = 1, \ldots, p$ is solution of the following optimisation problem

$$\min\left(\sum_{j=1}^{m} d_{ij}^{(t)} u_{ij}^{d}\right) \tag{27}$$

subject to

$$\sum_{j=1}^{m} u_{ij} = 1 \tag{28}$$

Therefore, we have

$$\sum_{j=1}^{m} d_{ij}^{(t)}(u_{ij}^{(t+1)})^d \leq \sum_{j=1}^{m} d_{ij}^{(t)}(u_{ij}^{(t)})^d \tag{29}$$

It means that

$$\begin{aligned}&\sum_{j=1}^{m}\left(u_{ij}^{(t+1)}\right)^d\left(\left\|\phi(x_i) - c_j^{(t)}\right\|^2 - \left(R_j^{(t)}\right)^2\right) \\ &\leq \sum_{j=1}^{m}\left(u_{ij}^{(t)}\right)^d\left(\left\|\phi(x_i) - c_j^{(t)}\right\|^2 - \left(R_j^{(t)}\right)^2\right) \leq \xi_i^{(t)}\end{aligned} \tag{30}$$

or

$$\sum_{j=1}^{m}\left(u_{ij}^{(t+1)}\right)^d\left\|\phi(x_i) - c_j^{(t)}\right\|^2 \leq \sum_{j=1}^{m}\left(u_{ij}^{(t+1)}\right)^d\left(R_j^{(t)}\right)^2 + \xi_i^{(t)} \tag{31}$$

It is certain that for $i = p+1, \ldots, n$ and $j = 1, \ldots, m$ we have

$$\left\|\phi(x_i) - c_j^{(t)}\right\|^2 \geq \left(R_j^{(t)}\right)^2 - \xi_{ij}^{(t)} \tag{32}$$

Hence, ($R^{(t)}$, $c^{(t)}$, $\xi_i^{(t)}$, $\xi_{ij}^{(t)}$, $U^{(t)}$) is feasible solution of optimisation problem (1) at time $t+1$. Since ($R^{(t+1)}$, $c^{(t+1)}$, $\xi_i^{(t+1)}$, $\xi_{ij}^{(t+1)}$, $U^{(t+1)}$) is minimal solution of this optimisation problem, Theorem 4 is proved.

3 Experiments

To show the performance of the proposed method, we established an experiment on 23 data sets in UCI repository as shown in Table 1. Most of them are two-class data sets and others are multi-class data sets. For each data set, we randomly selected one class and regarded its data samples as normal data samples. Data samples from the remaining class(es) were randomly selected to form a set of abnormal samples such that the ratio of normal samples and abnormal samples was kept to 12 : 1. We run cross validation with five folds and ten times and then take average of ten accuracies to obtain the final cross validation accuracy.

Table 1. Details of the data sets: #normal,#abnormal and d are number of normal, abnormal data and dimension of the input space, respectively

Datasets	*#normal*	*#abnormal*	*#d*
Astroparticle	2000	166	4
Australian	307	25	14
Bioinformatics	221	18	20
Breast Cancer	444	36	10
Diabetes	500	41	8
Dna	464	38	180
DelfPump	1124	93	64
Germany Number	300	24	24
Four class	307	25	2
Glass	70	5	9
Heart	164	13	13
Ionosphere	225	18	34
Letter	594	49	16
Liver Disorders	145	12	6
Sonar	97	8	60
Specf	254	21	44
Splice	517	43	60
SvmGuide1	2000	166	4
SvmGuide3	296	24	22
Thyroid	3679	93	21
Vehicle	212	17	18
Wine	59	5	13
USPS	1194	99	256

We compared the proposed method with SVDD [14] and HMS-SVDD [8]. The popular RBF Kernel $K(x, x') = e^{-\gamma\|x-x'\|^2}$ was applied and the parameter γ was varied in grid $\{2^i : i = 2j + 1, j = -8, \ldots, 2\}$. The parameter ν_1 and ν_2 were selected in grid $\{0.1, 0.2, 0.3, 0.4\}$. For FMS-SVDD, the number of spheres was chosen in grid $\{3, 5, 7, 9\}$ and parameter d was set to 1.5. To evaluate the classification rate, we employed the accuracy metric given by $acc = \frac{acc^+ + acc^-}{2}$ where

acc^+ and acc^- are the accuracies on positive (normal) and negative (abnormal) classes, respectively.

For most of the data sets, especially for the large data sets, the proposed method outperforms other kernel methods.

Table 2. Experimental results on 23 data sets in UCI repository

Datasets	*SVDD*	*HMS-SVDD*	*FMS-SVDD*
Astroparticle	91%	94%	**96%**
Australian	82%	82%	82%
Bioinformatics	69%	82%	**84%**
Breast Cancer	95%	98%	**99%**
Diabetes	65%	**71%**	70%
Dna	81%	97%	**97%**
DelfPump	69%	74%	**74%**
Germany Number	68%	70%	**72%**
Four class	93%	96%	**97%**
Glass	82%	88%	**90%**
Heart	83%	**86%**	85%
Ionosphere	88%	91%	**94%**
Letter	90%	95%	**96%**
Liver Disorders	62%	71%	**72%**
Sonar	63%	69%	**71%**
Specf	70%	**76%**	**76%**
Splice	57%	63%	**64%**
SvmGuide1	92%	**98%**	**98%**
SvmGuide3	63%	68%	**70%**
Thyroid	88%	92%	**94%**
Vehicle	58%	59%	**60%**
Wine	97%	**98%**	**98%**
USPS	93%	**95%**	94%

4 Conclusion

In this paper, we have presented a fuzzy approach to Multi-sphere Support Vector Data Description to provide a better description to data sets with mixture of distinctive distributions. Each sample is assigned a fuzzy membership function representing the degree of belonging of that sample to a hypersphere. We have theoretically proved that *structural risk* becomes smaller across iterations in the learning process. Experiments on 23 real data sets in UCI repository have showed a better performance of the proposed method in comparison to HMS-SVDD and SVDD.

References

[1] Ben-Hur, A., Horn, D., Siegelmann, H.T., Vapnik, V.: Support vector clustering. Journal of Machine Learning Research 2, 125–137 (2001)
[2] Boser, B.E., Guyon, I.M., Vapnik, V.: A training algorithm for optimal margin classifiers. In: Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory, pp. 144–152. ACM Press (1992)
[3] Boyd, S., Vandenberghe, L.: Convex Optimisation. Cambridge University Press (2004)
[4] Burges, C.J.C.: A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery 2, 121–167 (1998)
[5] Chen, Y., Zhou, X., Huang, T.S.: One-class svm for learning in image retrieval. In: ICIP (2001)
[6] Chiang, J.-H., Hao, P.-Y.: A new kernel-based fuzzy clustering approach:support vector clustering with cell growing. IEEE Transactions on Fuzzy Systems 11(4), 518–527 (2003)
[7] Le, T., Tran, D., Ma, W., Sharma, D.: An optimal sphere and two large margins approach for novelty detection. In: The 2010 International Joint Conference on Neural Networks (IJCNN), pp. 1–6 (2010)
[8] Le, T., Tran, D., Ma, W., Sharma, D.: A theoretical framework for multi-sphere support vector data description. In: Wong, K.W., Mendis, B.S.U., Bouzerdoum, A. (eds.) ICONIP 2010, Part II. LNCS, vol. 6444, pp. 132–142. Springer, Heidelberg (2010)
[9] Lee, K., Kim, W., Lee, K.H., Lee, D.: Density-induced support vector data description. IEEE Transactions on Neural Networks 18(1), 284–289 (2007)
[10] GhasemiGol, M., Monsefi, R., Yazdi, H.S.: Ellipse support vector data description. In: Palmer-Brown, D., Draganova, C., Pimenidis, E., Mouratidis, H. (eds.) EANN 2009. CCIS, vol. 43, pp. 257–268. Springer, Heidelberg (2009)
[11] Moya, M.M., Koch, M.W., Hostetler, L.D.: One-class classifier networks for target recognition applications, pp. 797–801 (1991)
[12] Scott, C.D., Nowak, R.D.: Learning minimum volume sets. Journal of Machine Learning Research 7, 665–704 (2006)
[13] Tax, D.M.J., Duin, R.P.W.: Support vector data description. Journal of Machine Learning Research 54(1), 45–66 (2004)
[14] Tax, D.M.J., Duin, R.P.W.: Support vector domain description. Pattern Recognition Letters 20, 1191–1199 (1999)
[15] Wu, M., Ye, J.: A small sphere and large margin approach for novelty detection using training data with outliers. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(11), 2088–2092 (2009)
[16] Xiao, Y., Liu, B., Cao, L., Wu, X., Zhang, C., Hao, Z., Yang, F., Cao, J.: Multi-sphere support vector data description for outliers detection on multi-distribution data. In: ICDM Workshops, pp. 82–87 (2009)