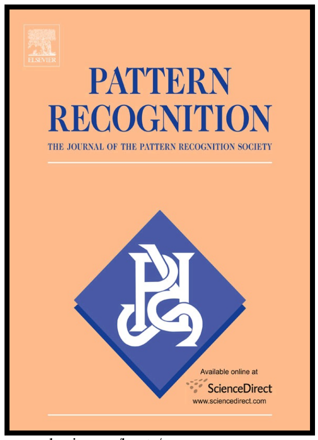


www.elsevier.com/locate/pr

Selective ensemble of SVDDs with Renyi entropy based diversity measure

Hong-Jie Xing, Xi-Zhao Wang

PII: S0031-3203(16)30195-9
DOI: http://dx.doi.org/10.1016/j.patcog.2016.07.038
Reference: PR5824

To appear in: *Pattern Recognition*

Received date: 22 August 2015
Revised date: 25 May 2016
Accepted date: 25 July 2016

Cite this article as: Hong-Jie Xing and Xi-Zhao Wang, Selective ensemble o SVDDs with Renyi entropy based diversity measure, *Pattern Recognitior* http://dx.doi.org/10.1016/j.patcog.2016.07.038

Selective ensemble of SVDDs with Renyi entropy based diversity measure

Hong-Jie Xing [a,*], Xi-Zhao Wang [b],

[a] *Key Laboratory of Machine Learning and Computational Intelligence, College of Mathematics and Information Science, Hebei University, Baoding 071002, China*

[b] *College of Computer Science and Software, Shenzhen Univsrsity, Shenzhen 518060, China*

Abstract

In this paper, a novel selective ensemble strategy for support vector data description (SVDD) using the Renyi entropy based diversity measure is proposed to deal with the problem of one-class classification. In order to obtain compact classification boundary, the radius of ensemble is defined as the inner product of the vector of combination weights and the vector of the radii of SVDDs. To make the center of ensemble achieve the optimal position, the Renyi entropy of the kernelized distances between the images of samples and the center of ensemble in the high-dimensional feature space is defined as the diversity measure. Moreover, to fulfill the selective ensemble, an ℓ_1-norm based regularization term is introduced into the objective function of the proposed ensemble. The optimal combination weights can be iteratively obtained by the half-quadratic optimization technique. Experimental results on two synthetic data sets and twenty benchmark data sets demonstrate that the proposed selective ensemble method is superior to the single SVDD and the other four related ensemble approaches.

Key words: One-class classification, Selective ensemble, Support vector data description, Renyi entropy

1 Introduction

One-class classification [1–3] is regarded as a machine learning task between supervised learning and unsupervised learning. It can efficiently deal with the problem of extreme class imbalance. In the training phase, only the samples in

* Corresponding author. E-mail address: hjxing@hbu.edu.cn

one-class can be used to train a classifier. Moreover, the testing samples can be classified as normal or novel by the trained classifier. There are many examples of one-class classification in our real world, such as machine fault detection [4], network intrusion detection [5], medical diagnosis [6], credit scoring [7], among others [8,9].

Support vector data description (SVDD) [10] is a generally used method as a one-class classifier. It establishes a hyper-sphere in the form of kernel expansion to distinguish the normal data from the novel data. The kernel function in the decision function maps the samples from the original space into a high-dimensional feature space while the explicit form of the mapping is not needed according to the 'kernel trick' [11]. When certain conditions are satisfied, SVDD is proved to be equivalent to one-class support vector machine (OCSVM) [12,10].

To make one-class classifier achieve more compact classification boundary, Tax and Duin [13] proposed the ensemble of one-class classifiers. They found that the ensemble can obviously improve the classification performance of one-class classifier. Seguí et al. [14] proposed the weighted bagging based ensemble of one-class classifiers. They utilized minimum spanning tree class descriptor as base classifiers. Zhang et al. [15] used locality preserving projection to reduce the dimensionality of the original data, trained several SVDDs upon the reduced data, and combined the outputs of the trained SVDDs. Hamdi and Bennani [16] proposed an ensemble of one-class classifiers by utilizing the orthogonal projection operator and the bootstrap strategy. Wilk and Woźniak [17] constructed the ensemble of one-class classifiers by fuzzy combiner. They utilized fuzzy rule based classifier as base classifier, while used fuzzy error correcting output codes and fuzzy decision templates as ensemble strategies. For tackling malware detection, Liu et al. [19] constructed random subspace method based ensemble of cost-sensitive twin one-class classifiers. Casale et al. [20] proposed the approximate polytope based ensemble of one-class classifiers. The methodology uses the geometrical concept of convex hull to define the boundary of the normal class, whilst utilizes random projections and ensemble decision process to judge whether a sample belongs to the convex hull in high-dimensional spaces. Furthermore, a tilling strategy was proposed to model non-convex structures. Krawczyk et al. [18] proposed the clustering-based ensemble of one-class classifiers. The clustering algorithm is utilized to split the whole normal class into the disjointed sub-regions. On each subregion, a single one-class classifier is trained. Finally, the outputs of all the one-class classifiers are combined together. Aghdam et al. [33] developed a new one-class classification method that can be trained with or without novel data and it can model the observation domain utilizing any binary classification approach. To mine data steams with concept drift, Czarnowski and Jedrzejowicz [34] proposed an instance selection and chunk updating based ensemble of one-class classifiers. Experimental results demonstrate that their method can outperform the well-known approaches for data streams with concept drift.

For the scenarios of two-class classification and multi-class classification, diversity is regarded as a key issue in classifier ensemble. Dietterich [21] compared the effectiveness of three ensemble methods, i.e., randomization, bagging, and boosting for improving the performance of the single decision tree. Through experiments he declared that randomization is competitive with bagging but not as accurate as boosting in the situation with little or no noise in the given training samples. Moreover, Dietterich also observed that the classifiers in the ensemble become less diverse as they become more accurate. Conversely, the classifiers become less accurate as they become more diverse. Kuncheva and Whitaker [22] studied ten measures of diversity between the base classifiers. They concluded that designing diverse classifiers is correct. However, in real-life pattern recognition problems, measuring diversity and utilizing the diversity to efficiently build better classifier ensemble is still an open problem. For the majority vote combiner, Brown and Kuncheva [23] first decomposed the classification error into three parts, i.e., individual accuracy, 'good' diversity, and 'bad' diversity. Moreover, they also declared that a larger value of the good diversity reduces the majority vote error, while a larger value of bad diversity increases the error. Recently, Sidhu et al. [24,25] studied the diversified ensemble approaches for the online stream data.

Similar to the cases of two-class classification and multi-class classification, diversity measure [26,27] acts an important role for the ensemble of one-class classifiers. Krawczyk and Woźniak [28,29] first investigated the diversity of ensemble for one-class classification and formulated five diversity measures. Moreover, Krawczyk and Woźniak [30] studied the relationship between the accuracy and diversity towards the ensemble of one-class classifiers. They proposed a novel ensemble strategy for one-class classification by assuring both high accuracy of individual one-class classifiers and high diversity among these classifiers. Besides the accuracy of individual one-class classifiers and the diversity of ensemble, the combination strategy also affects the performance of the ensemble of one-class classifiers. Menahem et al. summarized the commonly used combination rules and provided a list in literature [31]. However, these combination rules all rely on the estimated probability of sample given the normal class. In the study, the LSE (lease squares estimation)-based weighting [32] is utilized to directly combine the outputs of the decision functions of individual SVDDs.

As aforementioned, the classification boundary of SVDD in the high-dimensional feature space is hyper-sphere. After combined by the LSE-based weighting rule, the boundary of ensemble of SVDDs in the feature space is also a hyper-sphere. Fig. 1 illustrates an ensemble of SVDDs. It can be deduced from Fig. 1 that the performance of the ensemble of SVDDs is determined by its length of radius and location of center. Therefore, the study focuses on finding the optimal radius and center of ensemble rather than the highest diversity of ensemble.

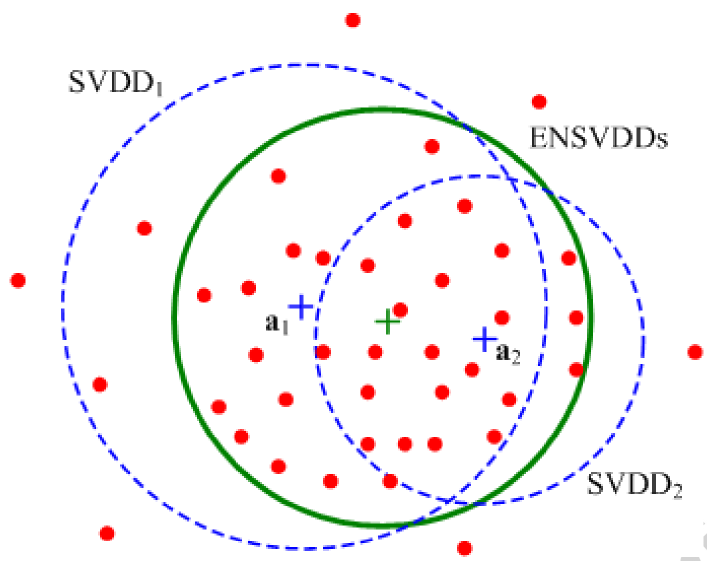


Fig. 1. Schematic diagram of the ensemble of SVDDs. $\mathbf{a}_1$ and $\mathbf{a}_2$ are the centers of the two SVDDs, while ENSVDDs is the ensemble of the two SVDDs.

Moreover, although an ensemble of classifiers often achieves better performance than one single classifier, the computational cost for obtaining the ensemble of these classifiers will become expensive when the number of base classifiers is large. To overcome the aforementioned disadvantage, Zhou et al. proved in [35] that it is better to ensemble part of the base classifiers rather than all of them. Li and Zhou [36] proposed a selective ensemble algorithm based on the regularization framework. Through solving a quadratic programming, they get the sparse solution of the vector of combination weights and implement the selective ensemble. Zhang and Zhou [37] proposed a linear programming based sparse ensemble method. Yan et al. [38] proposed a selective neural network ensemble classification algorithm for the incomplete data. It is noted in ensemble learning because the handling of uncertainty plays a key role for classifier performance improvement (e.g. ([39],[40]) and the selection of base classifier is very sensitive to the overall performance in bio-informatics ([41],[42]). Nevertheless, the existing selective ensemble approaches mainly concentrate on the supervised learning. Till now, there are too few work upon the selective ensemble of one-class classifiers. Krawczyk and Woźniak [43–46] investigated this issue and proposed four pruning strategies, i.e., multi-objective ensemble pruning, dynamic classifier selection method, firefly algorithm based ensemble pruning, and clustering-based pruning. Experimental results demonstrate that their methods outperform the state-of-the-art algorithms for selecting one-class classifiers from the given classifier committees. Parhizkar and Abadi [47] utilized a modified binary artificial bee colony algorithm to prune the ensemble of one-class classifiers and used the ordered weighted averaging operator to combine the outputs of base classifiers in the pruned ensemble.

In this study, we propose a selective ensemble strategy for SVDD to get the optimal combination weights of base classifiers. The proposed ensemble is mainly based on the Renyi entropy based diversity measure. The main contributions of the present study are as follows:

- The radius of ensemble is defined to be the inner product between the vector of combination weights and the vector of the radii of SVDDs. Therefore, minimizing the radius of ensemble can make the classification boundary of the ensemble of SVDDs as compact as possible.
- The Renyi entropy of the distance variable obtained by the kernelized distances between the images of samples and the center of ensemble in the feature space is defined as the diversity measure. Maximizing the Renyi entropy based diversity can make the center of ensemble attain the optimal position in the feature space.
- An ℓ_1-norm based regularization term of the vector of combination weights is introduced into the objective function of the proposed ensemble. Maximizing the ℓ_1-norm based regularization term can effectively fulfill the selective ensemble.

The rest of the paper is organized as follows. SVDD and Renyi entropy are briefly reviewed in Section 2. In Section 3, the proposed selective ensemble based on the Renyi entropy based diversity measure is expatiated. Experiments to validate the proposed ensemble method are conducted in Section 4. Finally, Section 5 concludes the study.

2 Preliminaries

In this section, the optimization problems and decision formula of SVDD are briefly introduced, while the mathematical expression and estimation calculation of Renyi entropy are briefly reviewed.

2.1 SVDD

SVDD was proposed by Tax and Duin [10]. It finds the smallest sphere enclosing all the normal data. Given N normal data $\{\mathbf{x}_i\}_{i=1}^N$ with $\mathbf{x}_i \in \mathcal{R}^d$, the original optimization problem of SVDD is given by

$$\begin{aligned}
&\min_{R,\mathbf{a},\boldsymbol{\xi}} \; R^2 + C\sum_{i=1}^{N}\xi_i \\
&s.t. \; \|\mathbf{x}_i - \mathbf{a}\|^2 \le R^2 + \xi_i, \; i = 1, 2, \ldots, N \\
&\quad\;\; \xi_i \ge 0, \; i = 1, 2, \ldots, N,
\end{aligned} \tag{1}$$

where R is the radius of the enclosing sphere, C is the trade-off parameter, ξ_i is the slack variable, and $\mathbf{a}$ is the center of the enclosing sphere. The optimization problem (1) can be solved by the Lagrange multiplier method. Moreover, substituting the inner products in the dual optimization problem of (1) by kernel functions, we can obtain the following dual optimization problem with nonlinear kernels

$$\begin{aligned}
&\min_{\boldsymbol{\alpha}} \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K(\mathbf{x}_i, \mathbf{x}_j) - \sum_{i=1}^{N}\alpha_i K(\mathbf{x}_i, \mathbf{x}_i) \\
&s.t. \; \sum_{i=1}^{N}\alpha_i = 1 \\
&\quad\;\; 0 < \alpha_i < C, \; i = 1, 2, \ldots, N.
\end{aligned} \tag{2}$$

For the choice of kernel functions $K(\cdot,\cdot)$, one can refer to literature [48].

Given a test sample $\mathbf{x}$, it can be classified as the normal data if the following condition holds

$$\|\phi(\mathbf{x}) - \mathbf{a}\| = \sqrt{K(\mathbf{x}, \mathbf{x}) - 2\sum_{i=1}^{N}\alpha_i K(\mathbf{x}, \mathbf{x}_i) + \sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j K(\mathbf{x}_i, \mathbf{x}_j)} \le R, \tag{3}$$

where $\phi(\cdot)$ is a nonlinear mapping function that maps the given samples from the original space into the high-dimensional feature space. Moreover, if (3) does not satisfy, $\mathbf{x}$ is classified as the novel data.

2.2 Renyi Entropy

For a statistic variable X with probability density function $f(X)$, its Renyi entropy is defined as

$$H_R(X) = \frac{1}{1-\alpha}\log\int f^{\alpha}(X)dX, \alpha > 0, \alpha \ne 1. \tag{4}$$

Especially, if $\alpha = 2$, (4) becomes

$$H_{R_2}(X) = -\log \int f^2(X)dX, \tag{5}$$

which is known as Renyi quadratic entropy.

Let $\{\mathbf{x}_i\}_{i=1}^N$ be a set of independent identical distribution samples with d features drawn from the probability density function $f(X)$. Therefore, $f(X)$ can be estimated by the Parzen window estimator, that is,

$$\hat{f}(X) = \frac{1}{N}\sum_{i=1}^{N} W_{\sigma^2}(\mathbf{x} - \mathbf{x}_i), \tag{6}$$

where $W_{\sigma^2}(\cdot)$ is the Parzen window and the scale parameter σ^2 controls its width. Typically, Gaussian kernel function is commonly chosen as the kernel function of the Parzen window.

$$W_{\sigma^2}(\mathbf{x} - \mathbf{x}_i) = G(\mathbf{x} - \mathbf{x}_i, \sigma^2) = \frac{1}{(2\pi)^{d/2}\sigma^d} \exp\left\{-\frac{(\mathbf{x} - \mathbf{x}_i)^T(\mathbf{x} - \mathbf{x}_i)}{2\sigma^2}\right\}. \tag{7}$$

According to the convolution theorem for Gaussian [49], we have

$$\int G(\mathbf{x} - \mathbf{x}_i, \sigma^2)G(\mathbf{x} - \mathbf{x}_j, \sigma^2)d\mathbf{x} = G(\mathbf{x}_i - \mathbf{x}_j, 2\sigma^2). \tag{8}$$

Utilizing (8), the information potential $\hat{V}_{R_2}(X)$ of (5) can be expressed as

$$\begin{aligned}
\hat{V}_{R_2}(X) &= \int \hat{f}^2(X)dX \\
&= \int \frac{1}{N}\sum_{i=1}^{N} G(\mathbf{x} - \mathbf{x}_i, \sigma^2)\frac{1}{N}\sum_{j=1}^{N} G(\mathbf{x} - \mathbf{x}_j, \sigma^2)d\mathbf{x} \\
&= \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\mathbf{x}_i - \mathbf{x}_j, 2\sigma^2).
\end{aligned} \tag{9}$$

Hence, the Renyi entropy of $\{\mathbf{x}_i\}_{i=1}^N$ can be obtained as

$$H_{R_2}(\{\mathbf{x}_i\}_{i=1}^N) = -\log \hat{V}_{R_2}(\{\mathbf{x}_i\}_{i=1}^N) = -\log\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G(\mathbf{x}_i - \mathbf{x}_j, 2\sigma^2)\right]. \tag{10}$$

For the Renyi entropy in (10), there are two issues should be declared. First, because the Renyi quadratic entropy is a lower bound of Shannon's entropy, it might be more efficient than Shannon's entropy for entropy maximization (cf. [50] p. 54). Second, it turns out that the Renyi entropy in (10) provides significant computational saving [51].

3 Selective Ensemble of SVDDs

3.1 Problem Formulation

Definition 1 *For M SVDDs, the radius of their ensemble is defined as*

$$\bar{r} = \sum_{k=1}^{M} w_k r_k = \mathbf{w}^T \mathbf{r}, \tag{11}$$

where r_k denotes the radius of the kth SVDD, $\mathbf{w} = (w_1, w_2, \ldots, w_M)^T$ is the vector of combination weights for the M radii, and $\mathbf{r} = (r_1, r_2, \ldots, r_M)^T$ is the vector consisting of the M radii.

To make the classification boundary of the ensemble of SVDDs as compact as possible, the square of its radius $\bar{r}^2$ needs to be minimized. However, to avoid SVDDs in ensemble obtain the same radius and center, part of the whole training samples are randomly selected with replacement to train each SVDD. Therefore, the initialization strategy of the proposed ensemble method is same with bagging. According to (11), $\bar{r}^2$ can be expressed as

$$\bar{r}^2 = \left(\mathbf{w}^T \mathbf{r}\right)^2 = \mathbf{w}^T \mathbf{r}\mathbf{r}^T \mathbf{w}. \tag{12}$$

Let the center vector of the kth SVDD be $\mathbf{a}_k$. The kernelized distance between the image of the ith sample $\mathbf{x}_i$ and $\mathbf{a}_k$ in the feature space is given by

$$d_{ik} = \|\phi(\mathbf{x}_i) - \mathbf{a}_k\| = \sqrt{K(\mathbf{x}_i, \mathbf{x}_i) - 2\sum_{j=1}^{N} \alpha_{kj} K(\mathbf{x}_i, \mathbf{x}_j) + \sum_{j=1}^{N}\sum_{l=1}^{N} \alpha_{kj}\alpha_{kl} K(\mathbf{x}_j, \mathbf{x}_l)}. \tag{13}$$

The different kernelized distances between the image of $\mathbf{x}_i$ and the centers of two different SVDDs in the feature space are illustrated in Fig.2.

Definition 2 *Let d_{ik} denote the kernelized distance between the image of the ith sample $\mathbf{x}_i$ and the center of the kth SVDD in the feature space. The kernelized distance between the image of $\mathbf{x}_i$ and the center of ensemble in the feature*

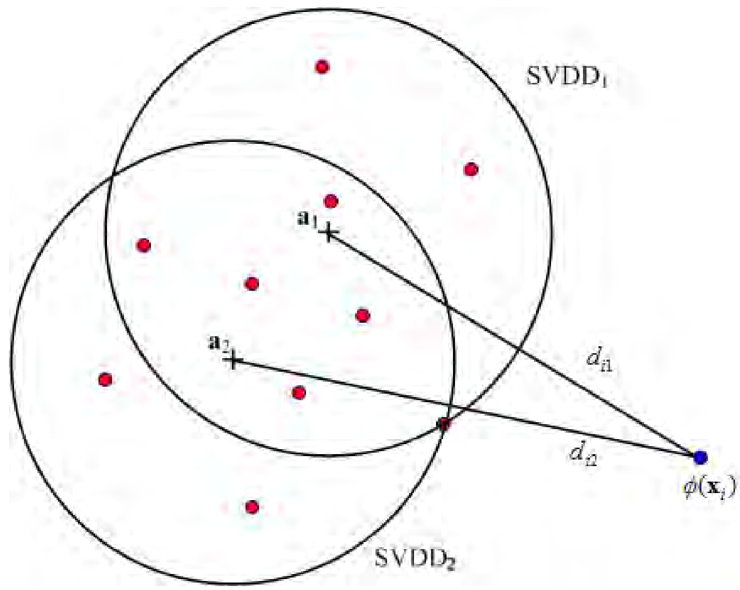


Fig. 2. The kernelized distance between the image of $\mathbf{x}_i$ and the centers of two different SVDDs in the feature space.

space is defined as

$$\bar{d}_i = \sum_{k=1}^{M} w_k d_{ik} = \mathbf{w}^T \mathbf{d}_i, \tag{14}$$

where $\mathbf{d}_i = (d_{i1}, d_{i2}, \ldots, d_{iM})^T$.

For the above two definitions, there are three issues should be mentioned as follows.

- Once the training procedure of the M SVDDs completes, they keep fixed. Therefore, the centers $\{\mathbf{a}_k\}_{k=1}^M$ and radii $\{r_k\}_{k=1}^M$ of the M SVDDs remain unchaged during the procedure for constructing an ensemble.
- The explicit expression of the center of ensemble is not given. One can deduce from (14) that the position of the center of ensemble moves as the distances $\{\bar{d}_i\}_{i=1}^N$ alter.
- According to the formulae (11) and (14), the location of the center and the length of the radius for an ensemble are both affected by tuning the vector of combination weights $\mathbf{w}$.

Definition 3 *The diversity of the distance variable* $\bar{D} = \{\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_N\}$ *can* *be measured by its Renyi entropy*

$$H_{R_2}(\bar{D}) = -\log \hat{V}_{R_2}(\bar{D}) = -\log \left[\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G\left(\bar{d}_i - \bar{d}_j, 2\sigma^2\right) \right]. \tag{15}$$

Note that the proposed diversity measure is used for measuring the diversity of the N distances $\bar{d}_1, \bar{d}_2, \ldots, \bar{d}_N$ rather than the diversity of the N SVDDs. According to (15), the high degree of scatter for the N distances can be obtained by maximizing the proposed diversity measure. Moreover, as the aforementioned issue, the position of the center of ensemble can be changed by altering the N kernelized distances between the N samples and the center of ensemble. Therefore, the optimal location of the center of ensemble can also be obtained by maximizing $H_{R_2}(\bar{D})$ in (15). Because the logarithm function is strictly increasing, maximizing $H_{R_2}(\bar{D})$ is equivalent to minimizing its corresponding information potential $\hat{V}_{R_2}(\bar{D})$.

To fulfill the selective ensemble, an ℓ_1-norm based regularization term $\|\mathbf{w}\|_1$ can be introduced. Note that the elements in the combination weight $\mathbf{w}$ are all no less than zero, i.e., $w_k \geq 0$ $(k = 1, 2, \ldots, M)$.

In summary, the optimization problem of the selective ensemble of SVDDs using the Renyi entropy based diversity measure is given by

$$\begin{aligned} &\min \ \mathbf{w}^T \mathbf{r} \mathbf{r}^T \mathbf{w} + \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} G\left(\bar{d}_i - \bar{d}_j, 2\sigma^2\right) - \lambda \|\mathbf{w}\|_1 \\ &s.t. \ w_k \geq 0, k = 1, 2, \ldots, M \end{aligned} \tag{16}$$

3.2 Solution

The second term in the objective function of (16) can be reformulated as

$$\begin{aligned} &\frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \exp\left\{ -\frac{(\bar{d}_i - \bar{d}_j)^2}{4\sigma^2} \right\} = \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \exp\left\{ -\frac{(\mathbf{w}^T \mathbf{d}_i - \mathbf{w}^T \mathbf{d}_j)^2}{4\sigma^2} \right\} \\ &= \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} \exp\left\{ -\frac{\mathbf{w}^T (\mathbf{d}_i - \mathbf{d}_j)(\mathbf{d}_i - \mathbf{d}_j)^T \mathbf{w}}{4\sigma^2} \right\}. \end{aligned} \tag{17}$$

Because $\|\mathbf{w}\|_1 = \sum_{k=1}^{M} |w_k|$ and $w_k \geq 0$ $(k = 1, 2, \ldots, M)$, we can get that

$$\|\mathbf{w}\|_1 = \sum_{k=1}^{M} w_k = \mathbf{w}^T \mathbf{1}_M, \tag{18}$$

where $\mathbf{1}_M$ is an M-dimensional vector with its elements are all one. Therefore, substituting (17) and (18) into (16), we have

$$\min\ \mathbf{w}^T\mathbf{r}\mathbf{r}^T\mathbf{w} + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\exp\left\{-\frac{\mathbf{w}^T(\mathbf{d}_i-\mathbf{d}_j)(\mathbf{d}_i-\mathbf{d}_j)^T\mathbf{w}}{4\sigma^2}\right\} - \lambda\mathbf{w}^T\mathbf{1}_M. \quad (19)$$

There are many methods for solving the optimization problem (19) in the literature, e.g. half-quadratic optimization technique [52,53], expectation-maximization (EM) method [54], and gradient-based method [55,56]. In the study, the half-quadratic optimization technique is utilized. According to the theory of the convex conjugated function [53], we have

Proposition 1 *For* $G(z) = \exp\left\{-\frac{z^2}{2\sigma^2}\right\}$, *there exists a convex conjugated function* φ, *such that*

$$G(z) = \sup_{\alpha\in\mathcal{R}^-}\left(\alpha\frac{z^2}{2\sigma^2} - \varphi(\alpha)\right). \quad (20)$$

Moreover, for a fixed z, *the supremum is reached at* $\alpha = -G(z)$ *[52].*

According to Proposition 1, the objection function (19) can be augmented as

$$J(\mathbf{w},\mathbf{P}) = \max_{\mathbf{w},\mathbf{P}}\left\{-\mathbf{w}^T\mathbf{r}\mathbf{r}^T\mathbf{w} - \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\left[p_{ij}\frac{\mathbf{w}^T(\mathbf{d}_i-\mathbf{d}_j)(\mathbf{d}_i-\mathbf{d}_j)^T\mathbf{w}}{4\sigma^2} - \varphi(p_{ij})\right] + \lambda\mathbf{w}^T\mathbf{1}_M\right\}, \quad (21)$$

where $\mathbf{P} = (p_{ij})_{N\times N}$ stores the auxiliary variables in the half-quadratic optimization.

The local optimization solution of (21) can be iteratively calculated by

$$\hat{p_{ij}}^{\tau+1} = -\exp\left\{-\frac{(\mathbf{w}^\tau)^T(\mathbf{d}_i-\mathbf{d}_j)(\mathbf{d}_i-\mathbf{d}_j)^T\mathbf{w}^\tau}{4\sigma^2}\right\} \quad (22)$$

and

$$\begin{aligned}\mathbf{w}^{\tau+1} &= \arg\max_{\mathbf{w}}\left\{-\mathbf{w}^T\mathbf{r}\mathbf{r}^T\mathbf{w} - \mathbf{w}^T\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\frac{p_{ij}(\mathbf{d}_i-\mathbf{d}_j)(\mathbf{d}_i-\mathbf{d}_j)^T}{4N^2\sigma^2}\right]\mathbf{w} + \lambda\mathbf{w}^T\mathbf{1}_M\right\} + \text{const},\\ &= \arg\max_{\mathbf{w}}\left[-\mathbf{w}^T\left(\mathbf{r}\mathbf{r}^T + \frac{1}{4N^2\sigma^2}\mathbf{D}\mathbf{L}\mathbf{D}^T\right)\mathbf{w} + \lambda\mathbf{w}^T\mathbf{1}_M\right], \quad (23)\end{aligned}$$

where τ denotes the τth iteration, $\mathbf{D} = (\mathbf{d}_1,\mathbf{d}_2,\ldots,\mathbf{d}_N)$, $\mathbf{L} = \mathbf{P}-\mathbf{Q}$ is the Laplacian matrix with the main diagonal entries of the diagonal matrix $\mathbf{Q}_{ii} = \sum_{j=1}^{N} p_{ij}$.

The problem (23) can be solved by finding the saddle point of the objective function and by taking the derivative of the objective function with respect to $\mathbf{w}$. Therefore, the local optimal solution of (23) in the $(\tau+1)$th iteration is given by

$$\hat{\mathbf{w}}^{\tau+1} = \frac{\lambda}{2}\left(\mathbf{r}\mathbf{r}^T + \frac{\mathbf{DLD}^T}{4N^2\sigma^2}\right)^{-1}\mathbf{1}_M. \tag{24}$$

It should be mentioned here that the objective function $J(\mathbf{w}, \mathbf{P})$ in (21) converges after certain iterations, which can be summarized as follows:

Proposition 2 *The sequence $J(\mathbf{w}^\tau, \mathbf{P}^\tau)$, $\tau = 1, 2, \ldots$ generated by the iterations (22) and (23) converges.*

PROOF. *According to Proposition 1 and (23), we can find that $J(\mathbf{w}^\tau, \mathbf{P}^\tau) \leq J(\mathbf{w}^{\tau+1}, \mathbf{P}^\tau) \leq J(\mathbf{w}^{\tau+1}, \mathbf{P}^{\tau+1})$. Therefore, the sequence $\{J(\mathbf{w}^\tau, \mathbf{P}^\tau), \tau = 1, 2, \ldots\}$ is non-decreasing. Moreover, it was shown in [55] that correntropy is bounded. Thus, we know that $J(\mathbf{w}^\tau, \mathbf{P}^\tau)$ is bounded. Consequently, we verify that $\{J(\mathbf{w}^\tau, \mathbf{P}^\tau), \tau = 1, 2, \ldots\}$ converges.*

As stated in Section 1, SVDD is proved to be equivalent to OCSVM. However, the following remark should be mentioned.

Remark 1 *The base classifier of the proposed ensemble method is fixed to SVDD. Although SVDD and OCSVM are equivalent in certain conditions, OCSVM cannot be directly used to substitute SVDD in the proposed ensemble method.*

The equivalence between SVDD and OCSVM can be verified in two different ways. In the first way, the dual optimization problem of SVDD is proved to be equivalent to that of OCSVM when the Gaussian kernel function is utilized and the trade-off parameter C of SVDD is equal to the inverse of the product of the number of training samples N and the regularization parameter ν of OCSVM [12]. In the second way, the primal optimization problem of SVDD is equivalent to that of OCSVM as long as the following three conditions hold [10]: (i) For OCSVM, the norm of its normal vector $\|\mathbf{w}\|$ equals one. (ii) For SVDD, the training samples $\mathbf{x}_i$ $(i = 1, 2, \ldots, N)$ and the center vector $\mathbf{a}$ are all normalized. (iii) The intercept term ρ of OCSVM equals $2 - R^2$ of SVDD. Moreover, C of SVDD equals the inverse of the product of the number of training samples N and ν of OCSVM.

For the first way to derive the equivalence between SVDD and OCSVM, it can be easily find that there is no relationship between the radius R of SVDD and the intercept term ρ of OCSVM. Therefore, the radius of ensemble for OCSVM cannot be obtained. For the second way, although the relationship between R and ρ is provided, normalizing the center vector $\mathbf{a}$ of SVDD and limiting the norm of the normal vector $\mathbf{w}$ of OCSVM be one may greatly reduce the classification performances of the final obtained SVDD and OCSVM. Moreover, for OCSVM, the diversity measure (15) cannot be obtained because there is no the definition of center vector in its whole training procedure. Hence, OCSVM cannot be directly used to substitute SVDD as the base classifier of the proposed ensemble method.

3.3 Algorithm

The whole procedure of training the proposed selective ensemble of SVDDs is summarized in Algorithm 1. As mentioned in subsection 3.1, for the step 1 in Algorithm 1, part of the whole training samples are randomly chosen with replacement and used for training each of the M SVDDs. Once the training procedure of the M SVDDs completes, their centers and radii keep unchanged. Thereafter, the length of radius and the location of center of ensemble are adjusted by the proposed ensemble strategy. At the same time, the redundant SVDDs in the ensemble can be removed.

Algorithm 1 Selective ensemble of SVDDs

Input: Training samples $\{\mathbf{x}_i\}_{i=1}^N$, number of SVDDs M, maximum number of iterations I_{HQ}, regularization coefficient λ
Output: Optimal vector of combination weights $\mathbf{w}^*$, M trained SVDDs
Initialization: Width parameter of Gaussian kernel function γ, trade-off parameter C, width parameter of Renyi entropy function σ
Step 1: Train M SVDDs with the parameters γ and C.
Step 2: Randomly initialize the combination weights $w_k(k = 1, 2, \ldots, M)$ and ensure the summation of them equals one.
Step 3: Update the combination weights
for $\tau = 1, 2, \ldots, I_{HQ}$ **do**
  Update the auxiliary variables $p_{ij}(i, j = 1, 2, \ldots, N)$ by (22).
  Update the vector of combination weights $\mathbf{w}$ by (24).
end for

As is well known, the training complexity of SVDD is $O(N^3)$ [57]. Therefore, the computational cost of Step 1 in Algorithm 1 is $O(MN^3)$. For Step 3, the calculation of the kernelized distrance matrix $\mathbf{D}$ in each iteration needs $O(MN^3)$ operations. The computational complexity of the auxiliary matrix $\mathbf{P}$ in each iteration is $O\left(N^2(M^3 + M^2)\right)$. Moreover, the calculation of the vector of combination weights $\mathbf{w}$ in each iteration takes $O\left(MN^2 + (N+2)M^2 + M^3\right)$ operations according to (24). Therefore, the overall computational cost for Step 3 is $O\left(I_{HQ}[MN^3 + (M^3 + M^2 + M)N^2 + (N+2)M^2 + M^3]\right)$, where I_{HQ} is the number of the half-quadratic optimization. Finally, the total computational complexity of Algorithm 1 is $O\left((I_{HQ}+1)MN^3 + I_{HQ}[(M^3+M^2+M)N^2+(N+2)M^2+M^3]\right)$. Usually, $I_{HQ} \gg 1$ and $N \gg M$. The computational cost of Algorithm 1 is approximately $O(I_{HQ}(MN^3+M^3N^2))$.

Given a test sample $\mathbf{x}_{ts}$, the kernelized distances between its image and the M centers $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_M$ in the feature space are calculated as follows

$$d_{ts}^k = \|\phi(\mathbf{x}_{ts}) - \mathbf{a}_k\| = \sqrt{K(\mathbf{x}_{ts}, \mathbf{x}_{ts}) - 2\sum_{j=1}^{N} \alpha_{kj} K(\mathbf{x}_{ts}, \mathbf{x}_j) + \sum_{j=1}^{N}\sum_{l=1}^{N} \alpha_{kj}\alpha_{kl} K(\mathbf{x}_j, \mathbf{x}_l)}. \quad (25)$$

Therefore, the kernelized distance between the image of $\mathbf{x}_{ts}$ and the center of ensemble in the feature space is given by

$$\bar{d}_{ts} = \sum_{k=1}^{M} w_k^* d_{ts}^k. \quad (26)$$

Moreover, the radius of the ensemble is $\bar{r}^* = \sum_{k=1}^{M} w_k^* r_k$. Finally, the test sample $\mathbf{x}_{ts}$ can be classified as the normal data if $\bar{d}_{ts} \leq \bar{r}^*$ satisfies. Otherwise, $\mathbf{x}_{ts}$ is classified as the novel data.

4 Experimental Results

In the following experiments, the geometric mean (g-mean) is used to measure the performances of the different methods. The expression of g-mean is given by [58]

$$g = \sqrt{a^+ \times a^-}, \quad (27)$$

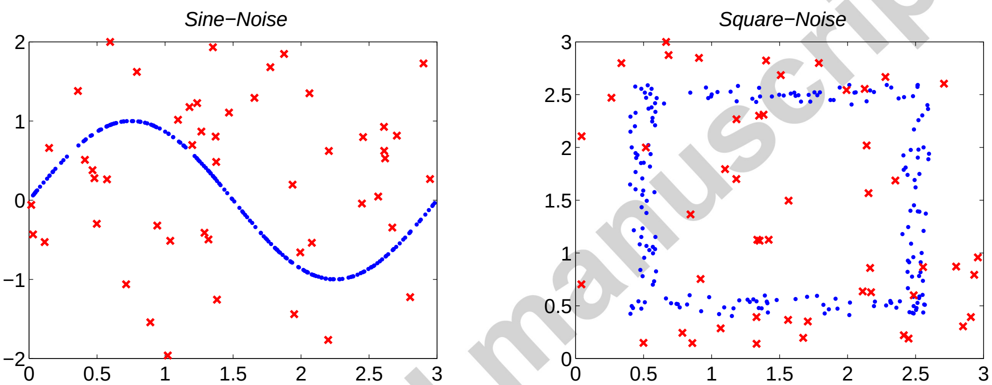
where a^+ and a^- denote the classification accuracy rates of a certain classifier upon the normal and novel data, respectively. For SVDD and the selective ensemble of SVDDs (SESVDDs), the Gaussian kernel function $K(\mathbf{x}, \mathbf{y}) = \exp\{-\gamma\|\mathbf{x} - \mathbf{y}\|^2\}$ are selected. For SESVDDs, its base classifiers are all constructed on the 80% samples randomly selected with replacement from each training set. During the course of training SESVDDs, the base classifiers with their combination weights satisfying $\frac{w_k}{\sum_{l=1}^{M} w_l} < \frac{1}{M}$ are discarded. In addition, all the codes are implemented in Matlab.

4.1 Synthetic Data Sets

To validate the effectiveness of SESVDDs, two synthetic data sets are generated. The number of base classifiers in SESVDDs and the maximum number of iterations for the half-quadratic optimization are both taken as 20. The description of the two synthetic data sets is as follows.

Sine-Noise: 200 noise-free samples are randomly chosen from the sine curve along $y = \sin\left(\frac{3}{2}\pi x\right)$ with $x \in [0, 3]$, while 50 noise are randomly distributed in the area $\{(x, y)|x \in [0, 3], y \in [-2, 2]\}$. Fig. 3(a) illustrates all the samples and noise.

Square-Noise: 200 noise-free samples are randomly selected in the square $\{(x, y)|x \in [0.4, 2.6], y \in [0.4, 0.6] \cup [2.4, 2.6]\} \cup \{(x, y)|x \in [0.3, 0.6] \cup [2.4, 2.6], y \in [0.4, 2.6]\}$, while 50 noise are randomly distributed in the area $\{(x, y)|x, y \in [0, 3]\}$. Fig. 3(b) shows all the samples and noise.

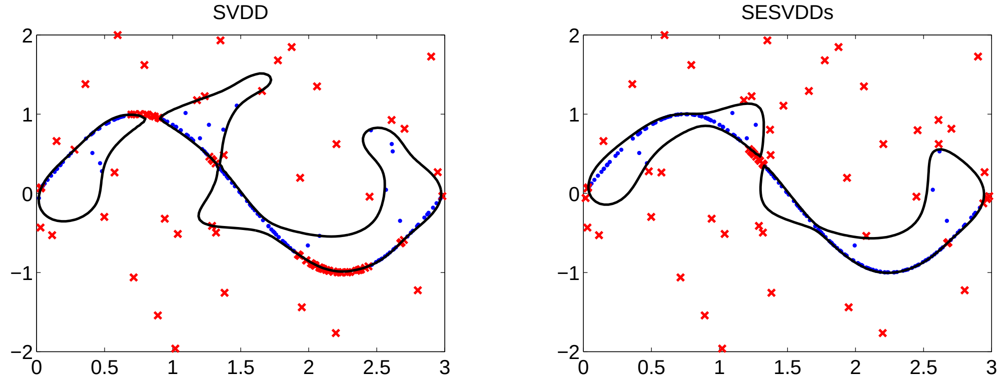


(a) (b)

Fig. 3. The two synthetic data set, where the samples with the blue dot are noise-free samples, while the samples with red cross are noise. (a) The *Sine-Noise* data set. (b) The *Square-Noise* data set.

For *Sine-Noise*, the parameter of the Gaussian kernel function and the trade-off parameter for SVDD are taken as $\gamma = 40$ and $C = 0.2$, respectively. The values of the parameters of the Gaussian kernel function and the trade-off parameter for SESVDDs are both the same with their counterparts of SVDD. Moreover, the width parameter of the Renyi entropy function and the regularization coefficient for SESVDDs are taken as $\sigma = 1$ and $\lambda = 1$, respectively. The results of the two methods are shown in Fig. 4. The number of base classifiers in the trained SESVDDs is 8.

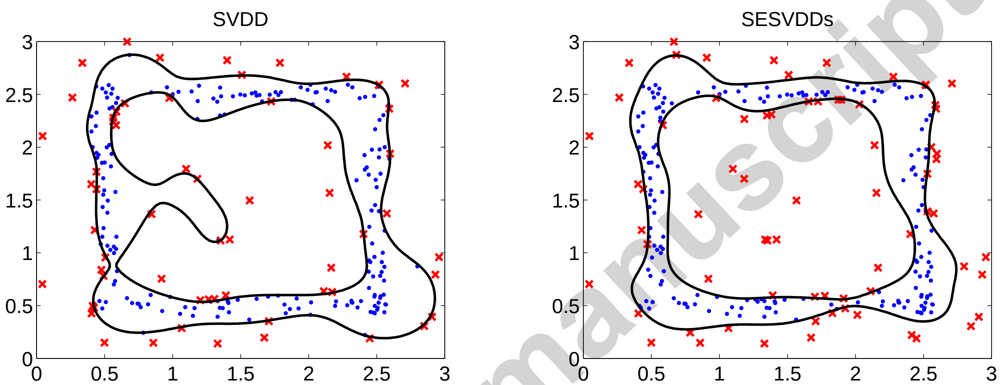
For *Square-Noise*, the settings of parameters for SVDD and SESVDDs are all the same with their corresponging roles upon *Sine-Noise*. The results of the two approaches are demonstrated in Fig. 5. The number of base classifiers in



(a) (b)

Fig. 4. The classification results of the two methods upon *Sine-Noise*. (a) SVDD with g-mean 0.6825. (b) SESVDDs with g-mean 0.8567.

the obtained SESVDDs is 11.



(a) (b)

Fig. 5. The classification results of the two methods upon *Square-Noise*. (a) SVDD with g-mean 0.7669. (b) SESVDDs with g-mean 0.8349.

According to the results demonstrated in Figs. 4 and 5 together with the g-mean of the two methods, one can easily find that the proposed SESVDDs is more robust against noise than SVDD upon the two synthetic data sets.

To observe the influence of the initial size of SESVDDs on its final classification performance, the initial number of SVDDs ranges from 10 to 300 with step length 10. The values of g-mean for SESVDDs upon the two synthetic data sets against the different numbers of SVDDs are shown in Fig. 6. For *Sine-Noise*, the optimal value of g-mean for SESVDDs is 0.9157 as the initial number of SVDDs is taken as 190, 200, or 210. One can easily observe from Fig. 6(a) that the value of g-mean for SESVDDs keeps unchanged as the initial number of SVDDs is bigger than 240. For *Square-Noise*, the optimal value of g-mean for SESVDDs is 0.8442 as the initial number of SVDDs is taken as 270. It can be observe from Fig. 6(b) that the values of g-mean for SESVDDs keeps fixed as the initial number of SVDDs is bigger than 280. Therefore, it can be deduced from Fig. 6 that the performance of SESVDDs becomes better as the initial number of SVDDs becomes larger. When the initial number of SVDDs attains certain value, SESVDDs can obtain the optimal performance. Finally, the performance of SESVDDs keeps unchanged as the number becomes larger enough.

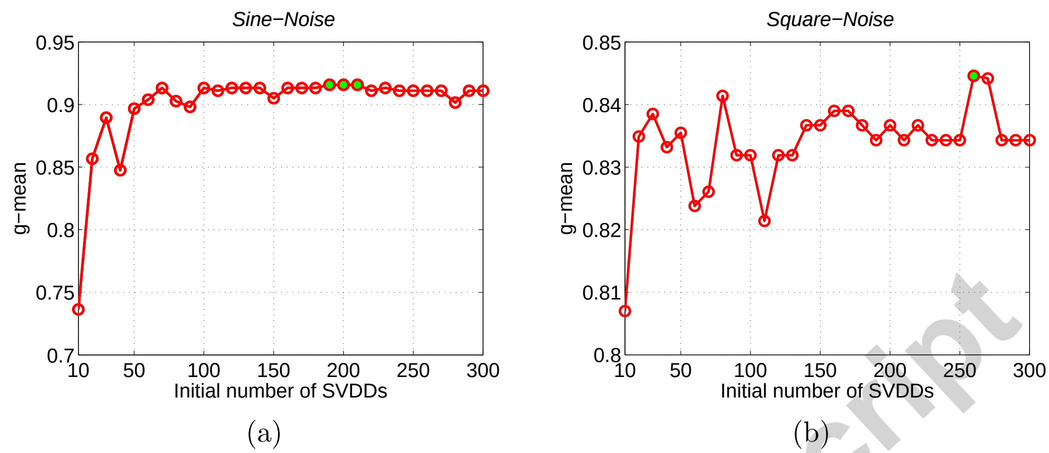


Fig. 6. The classification performance of SESVDDs with respect to the different initial number of SVDDs upon the two synthetic data sets. (a) The *Sine-Noise* data set. (b) The *Square-Noise* data set.

The final numbers of SVDDs corresponding to their initial number of SVDDs for SESVDDs upon the two synthetic data sets are illustrated in Fig. 7. One can find from Fig. 7 that the final numbers of SVDDs are approximately half of their corresponding initial number of SVDDs. Hence, SESVDDs can greatly reduce the number of SVDDs in its initial ensemble.

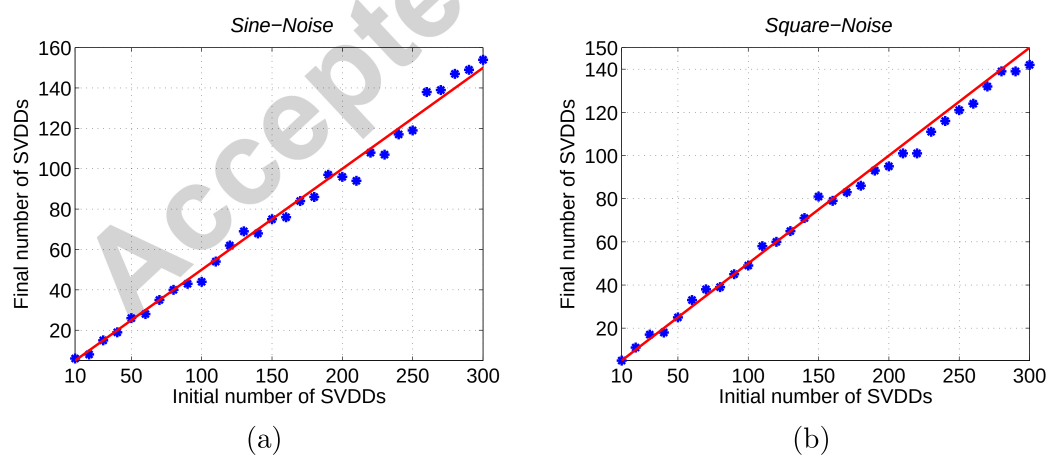


Fig. 7. The final numbers of SVDDs against the initial numbers of SVDDs for SESVDDs upon the two synthetic data sets. (a) The *Sine-Noise* data set. (b) The *Square-Noise* data set.

4.2 Benchmark Data Sets

To further validate the proposed ensemble method, it is compared with its related five approaches, i.e., single SVDD, ensemble of SVDDs by bagging [59] , ensemble of SVDDs by AdaBoost [60], random subspace method based ensemble of SVDDs (RSMESVDDs) [61], and clustering based ensemble of SVDDs (CESVDDs) [18] on the twenty benchmark data sets. Nine of the twenty benchmark data sets are chosen from the Rätsch's benchmark data sets [1] , while the rest eleven are taken from the UCI machine learning repository [62]. However, the above twenty benchmark data sets are initially used for two-class classification. To make them fit for one-class classification, the samples in one class of the given data set are used as normal data, and the samples in the other class are utilized as novel data. Furthermore, for each data set, its training set consists of 70% samples randomly chosen without replacement from the normal data, while the rest 30% samples of the normal data and the whole novel data are utilized for testing. The description of the twenty benchmark data sets is summarized in Table 1.

Table 1
The twenty benchmark data sets used in the experiments

Data sets	N_{normal}	N_{novel}	$N_{feature}$	N_{train}	N_{test}
Banana	2376	2924	2	1663	3637
Banknote Authentication	610	762	4	427	945
Blood Transfusion	178	570	4	125	623
Breast Cancer	77	186	9	54	209
Cancer	239	444	9	167	516
Cleverland Heart	214	83	13	150	147
Diabetis	268	500	8	188	580
Flare Solar	94	50	9	66	78
German	300	700	20	210	790
Heart	120	150	13	84	186
Hepatitis	123	32	19	86	69
Image	1188	898	18	832	1254
Liver	145	200	6	102	243
Parkinsons	147	48	22	103	92
Pima	268	500	8	188	580
Sonar	111	97	60	78	130
Twonorm	3703	3697	20	2592	4808
Waveform	1647	3353	21	1153	3847
Wdbc	212	357	9	148	421
Wholesale Customers	298	142	7	209	231

Note: N_{normal}–Number of normal data; N_{novel}–Number of novel data; $N_{feature}$–Number of features; N_{train}–Number of training data; N_{test}–Number of testing data.

[1] `http://theoval.cmp.uea.ac.uk/~gcc/matlab/default.html#benchmarks`

To make the single SVDD achieve better performance, its two parameters, i.e., the trade-off parameter C and the width parameter of the Gaussian kernel function γ are exhaustively searched within the domains $\{0.01, 0.025, 0.05, 0.075, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ and $\{0.0003, 0.0012, 0.005, 0.0078, 0.0312, 0.125, 0.5, 50, 5000\}$, respectively. The optimal values of C and γ for the single SVDD upon the twenty benchmark data sets are summarized in Table 2.

Table 2
The optimal parameters of single SVDD on the twenty benchmark data sets

Data sets	C	γ
Banana	0.1	5000
Banknote Authentication	0.025	0.125
Blood Transfusion	0.3	0.125
Breast Cancer	0.7	5000
Cancer	0.05	5000
Cleverland Heart	0.2	5000
Diabetis	0.4	50
Flare Solar	0.4	5000
German	0.5	5000
Heart	0.3	5000
Hepatitis	1	5000
Image	0.1	5000
Liver	0.2	0.0312
Parkinsons	0.2	5000
Pima	0.2	0.125
Sonar	0.05	50
Twonorm	0.1	5000
Waveform	0.1	5000
Wdbc	0.075	5000
Wholesale Customers	0.3	50

For the five ensemble approaches, namely, bagging, AdaBoost, RSMESVDDs, CESVDDs, and the proposed method, the parameters of their base classifiers are all assigned with the same values as those of the single SVDD on each benchmark data set. The number of base classifiers in the five ensemble methods are all taken as 50. For RSMESVDDs, the percentage of features retained in each training set is fixed at 75% and the trained SVDDs are combined by the majority voting rule. For CESVDDs, the PBMF-index based fuzzy c-means [63] is utilized to split the sample space. Moreover, the regularization coefficient of the proposed method, i.e. λ is taken as 1 in the following experiments. The maximum number of iterations for the half-quadratic optimization is taken as 20. The width parameter of the Renyi entropy function σ is chosen from $\{2^{-10}, 2^{-9}, \ldots, 2^{9}, 2^{10}\}$. Through experiments, we find that SESVDDs achieves better performance on all the twenty benchmark data set as $\sigma = 2^{10}$.

The average testing accuracy rates together with their corresponding standard deviations for the 20 trials of the six different approaches upon the twenty benchmark data sets are summarized in Table 3. Furthermore, the paired T-test and Wilcoxon rank-sum test are conducted to examine whether the performance improvement achieved by SESVDDs over the other five methods is statistically significant.

The results shown in Table 3 indicate that the proposed SESVDDs is statistically different from the other five methods, i.e., single SVDD, ensemble of SVDDs by bagging, ensemble of SVDDs by AdaBoost, random subspace method based ensemble of SVDDs, and clustering baded ensemble of SVDDs on all the twenty data sets except *Diabetis*, *Hepatitis*, and *Wholesale Customers*. Moreover, the generalization ability of SESVDDs is superior to the other five approaches. Taking the average testing accuracy rate into consideration, the values of standard deviation in Table 3 show that SESVDDs is more stable than the other five methods on all the twenty benchmark data sets.

For the results in Table 3, there are two issues to be mentioned as follows.

- In comparison with ensemble of SVDDs by bagging, ensemble of SVDDs by AdaBoost, random subspace method based ensemble of SVDDs, and clustering based ensemble of SVDDs, SESVDDs achieves better performance. The main reason lies that minimizing the weighted combination of radii of base classifiers can make the proposed ensemble method obtain the optimal length of radius, while maximizing the Renyi entropy based diversity upon the kernelized distances between the samples and the center of ensemble can make the proposed ensemble method obtain the optimal location of center.
- For SESVDDs, about half of base classifiers (25 of 50) are discarded on the twenty data sets. Therefore, introducing the ℓ_1-norm based regularization term into the objective function of SESVDDs can effectively get rid of the redundant base classifiers in ensemble.

In addition, the relationships between the classification performances of the five ensemble methods and the sizes of ensemble upon the four benchmark data sets are demonstrated in Fig. 8. The optimal values of g-mean for bagging upon the four benchmark data sets are 20, 160 (320, 340, or 400), 20, and 80 (or 100), respectively. The optimal values for AdaBoost are 80 (or 120), 60, 160, and 260 (or 400). The optimal values for RSMESVDDs are 200, 320, 80, and 80. The optimal values for SESVDDs are 280, 280 (320, or 340), 20 (or 60), and 320 (or 340). For each of the four data sets, the values of g-mean for CESVDDs are all the same. According to the above observations and the variation trends of the performance curves in Fig. 8, one can obtain the following outcomes.

- For bagging, its optimal classification performance achieves as the size of

Table 3. The average testing accuracy rates and the standard deviations of the six different methods on the twenty benchmark data sets (%)

Data sets	SVDD	Bagging	AdaBoost	RSMESVDDs	CESVDDs	SESV
	P_1, P_2	P_1, P_2	P_1, P_2	P_1, P_2	P_1, P_2	N_w
Banana	62.16±35.31	80.68±4.58	80.79±8.33	78.05±10.02	81.35±2.10	95.76±
	0.0224,6.80E-008	7.01E-007,6.80E-008	0.0002,4.88E-004	2.38E-013,6.48E-008	3.36E-026,2.85E-008	26
Banknote Authentication	87.44±0.27	88.03±0.30	87.63±0.28	85.47±0.22	87.20±0.15	88.18±
	1.48E-019,6.80E-008	5.11E-011,1.58E-008	3.30E-017,3.65E-008	4.75E-072,6.80E-008	1.101E-073,6.80E-008	24
Blood Transfusion	72.42±2.42	75.79±1.67	74.08±1.92	70.84±0.00	74.46±0.00	76.55±
	3.70E-013,1.15E-008	3.54E-005,3.65E-008	5.90E-010,1.15E-008	2.53E-036,6.80E-008	1.24E-017,6.80E-008	25
Breast Cancer	77.98±4.54	81.89±2.88	81.50±3.21	81.03±5.98	82.92*pm*2.86	86.09±
	0.0013,0.0028	0.0026,0.0052	0.0068,0.0068	1.55E-006,1.06E-007	1.98E-010,2.80E-008	26
Cancer	90.75±0.46	90.88±0.40	90.94±0.38	90.53±0.08	89.67±0.00	91.06±
	1.67E-007,1.55E-007	4.78E-008,3.65E-008	0.0003,0.0021	1.11E-009,1.15E-008	0,6.80E-008	25
Cleverland Heart	53.60±5.86	56.29±5.04	55.19±5.54	49.09±5.70	54.90±5.87	58.42±
	9.10E-009,1.31E-008	8.60E-009,1.31E-008	7.85E-008,1.27E-007	2.52E-008,1.58E-006	4.86E-010,3.16E-008	25
Diabetis	70.71±9.87	72.07±2.88	72.20±2.17	70.57±3.05	70.43±2.89	72.93±
	0.2187,0.4028	0.0220,0.0810	0.1783,0.3143	1.64E-005,2.69E-006	2.08E-006,3.94E-007	25
Flare Solar	60.74±7.84	61.18±6.92	61.48±5.87	66.33±5.46	56.60±5.58	70.06±
	0.0008,1.31E-005	0.0002,1.14E-004	0.0004, 1.28E-004	0.0122,1.19E-005	2.45E-010,3.28E-008	24
German	71.59±15.06	78.65±1.31	78.17±1.48	73.58±2.28	79.19±1.14	82.12±
	0.0062,1.32E-008	3.54E-009,2.21E-008	6.00E-007,3.42E-008	1.56E-006,4.18E-008	0.0042,1.31E-008	24
Heart	75.98±4.71	81.17±3.89	79.46±3.82	74.80±3.42	80.15±2.82	84.35±
	7.19E-11,1.42E-008	9.89E-009,1.66E-007	3.14E-009,4.45E-008	1.02E-016,6.75E-008	1.10E-012,6.80E-008	24
Hepatitis	64.83±18.15	64.39±8.28	66.71±10.22	65.08±9.86	66.82±9.92	67.60±
	0.6465,0.1075	0.0096,0.0085	0.7848,0.9246	0.0151,0.0090	0.5323,0.5250	24
Image	54.83±3.34	55.82±1.54	54.11±0.80	53.18±0.67	58.37±2.52	60.77±
	4.90E-008,5.60E-008	1.47E-012,1.67E-008	1.92E-13,1.48E-008	1.66E-021,6.80E-008	1.47E-008,6.79E-008	25
Liver	75.70±1.79	83.49±0.94	80.44±0.96	71.41±2.13	79.06±0.86	84.57±
	1.08E-008,6.80E-008	0.0002,0.0008	7.34E-008,4.16E-008	3.06E-025,6.56E-008	9.78E-024,5.79E-008	25
Parkinsons	52.51±5.16	55.76±5.87	54.02±5.74	53.08±6.54	52.04±4.60	57.96±
	4.16E-007,1.36E-008	0.0002,0.0144	1.27E-006,1.05E-006	6.99E-008,1.20E-006	6.14E-011,6.80E-008	25
Pima	72.63±2.78	79.92±1.94	77.27±1.95	70.50±4.52	78.61±1.97	80.81±
	9.77E-013,3.07E-008	4.98E-005,8.35E-005	7.65E-010,2.78E-008	4.88E-017,6.80E-008	2.18E-010,6.79E-008	26
Sonar	86.79±1.73	91.56±1.60	89.36±2.14	79.30±3.45	90.48±3.16	92.12±
	8.32E-007,3.24E-007	0.0402,0.020	5.86E-006,1.41E-006	4.95E-021,6.80E-008	0.0004,0.0009	25
Twonorm	87.83±7.72	90.74±1.11	91.10±1.39	89.65±4.87	91.08±1.23	96.17±
	0.0027,0.0088	1.95E-009,6.39E-008	1.98E-008,5.98E-008	1.28E-012,6.69E-008	1.84E-022,6.80E-008	25
Waveform	87.22±4.98	88.37±1.57	89.01±2.74	86.59±5.28	87.60±3.46	93.53±
	0.0002,0.0006	8.06E-012,5.32E-008	9.69E-006,4.10E-008	1.12E-010,1.67E-008	8.83E-015,6.79E-008	25
Wdbc	83.92±4.22	85.49±1.91	83.29±1.73	83.19±3.15	79.57±8.12	87.03±
	0.0003,0.0005	8.02E-012,6.91E-007	1.35E-015,2.88E-008	4.53E-012,6.80E-008	1.46E-007,1.05E-006	26
Wholesale Customers	74.24±2.72	75.35±2.41	74.22±3.48	70.71±4.57	61.67±8.20	77.97±
	0.1901,0.2931	0.2685,0.4291	0.2555,0.5108	0.0978,0.2125	0.1537,0.3460	23

Note: P_1–P-value for paired T-test; P_2–P-value for Wilcoxon rank-sum test; $N_{w_{nz}}$–Number of non-zero elements in the vector of combination weights.

ensemble is 20 for *Banknote Authentication* and *Pima*. However, the variation trends of the performance curves for the two cases are both very gently. For *Cancer* and *Wdbc*, the optimal performance of bagging is obtained as the size of ensemble is no less than 80. Overall, bagging may perform better with larger size of ensemble.

- For CESVDDs, increasing the number of SVDDs in its ensemble cannot obtain better classification performance. Hence, in order to reduce the training and testing costs, CESVDDs prefers smaller ensemble set-ups.
- For AdaBoost, RSMESVDDs, and SESVDDs, their optimal performances are all achieved as the size of ensemble is no less than 60. Therefore, The three ensemble approaches are all performing better with larger size of ensemble.
- Among the range [20,400] for the size of ensemble, SESVDDs achieves the optimal classification performance in comparison with the other four ensemble strategies upon the four data sets.

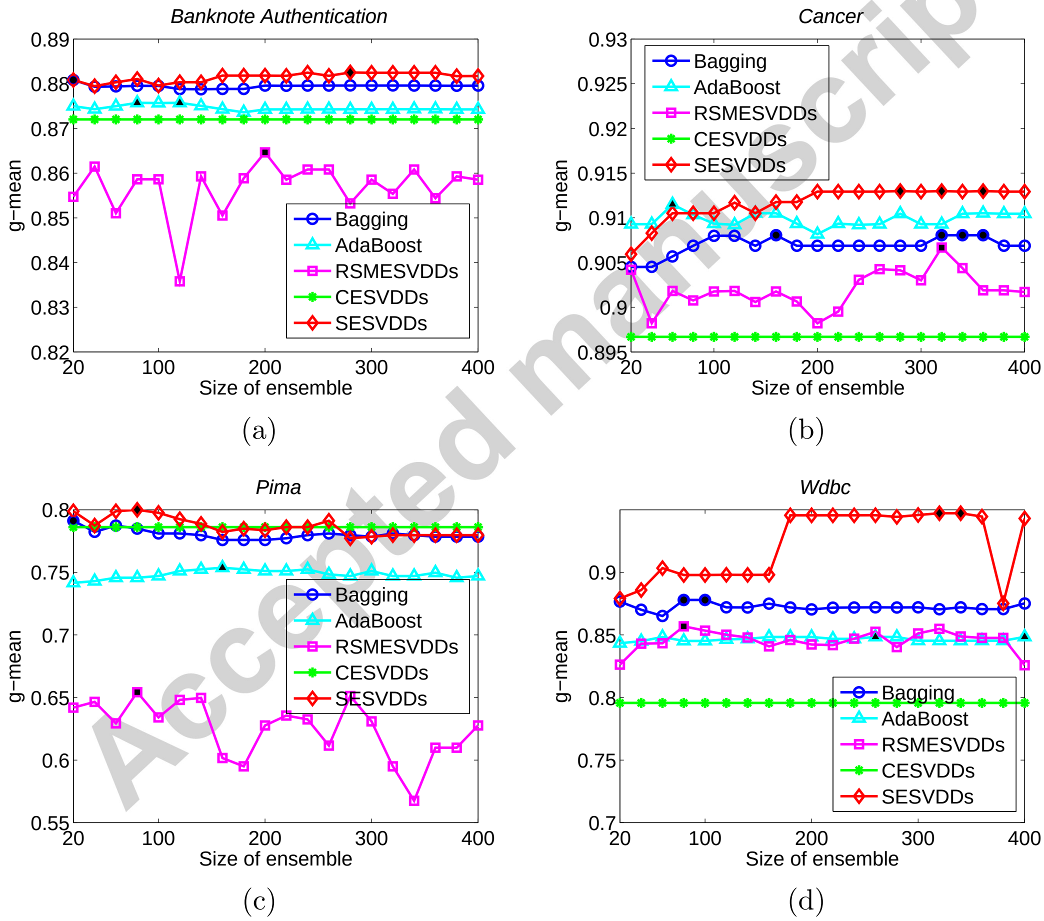


Fig. 8. The classification performances of the five different ensemble approaches with respect to the different numbers of SVDDs upon the four benchmark data sets. (a) *Banknote Authentication*. (b) *Cancer*. (c) *Pima*. (d) *Wdbc*.

5 Conclusions

To improve the generalization ability of SVDD, a novel selective ensemble strategy, named selective ensemble of SVDDs (SESVDDs), is presented. In SESVDDs, the radius of ensemble is defined and minimized to obtain the compact classification boundary. The diversity measure based on the Renyi entropy is proposed and maximized to get the optimal location of center of ensemble. Moreover, an ℓ_1-norm based regularization term is introduced into the objective function of SESVDDs to fulfill the selective ensemble. In comparison with the single SVDD and the other four ensemble approaches, SESVDDs demonstrates better anti-noise ability and generalization performance on the two synthetic and twenty benchmark data sets.

To make the proposed ensemble method more promising, there are four tasks for future investigation. First, it is a tough issue to find the appropriate width parameter σ for the Renyi entropy base diversity measure of SESVDDs. The heuristic methods for choosing the width parameter will be considered, such as the Silverman's rule [64]. Second, the classification performance and anti-noise ability of SESVDDs in difficult learning scenarios will be further examined. The training procedure of SESVDDs is too long when it is utilized to deal with large data sets. Therefore, it is necessary to find a more efficient training strategy for SESVDDs. Moreover, the anti-noise ability of SESVDDs on the synthetic data sets is verified. It will be further tested on the benchmark data sets. Third, one-class classifiers can be utilized to deal with the muti-class classification task [65–67]. However, through experiments we find that the computational cost of the proposed ensemble is very high when it is utilized to tackle the multi-class classification problems. The proposed ensemble strategy will be improved to make it fit for efficiently dealing with the multi-class classification task. Fourth, The classification performance of SESVDDs can be further improved by independently selecting different parameters for each SVDD in the ensemble rather than assigning the same parameter for all the SVDDs. However, the procedure of choosing the appropriate parameters for each SVDD is time-consuming, which may greatly increase the training cost of SESVDDs. In our future work, we will try to design a heuristic method for searching appropriate parameters for each SVDD in the ensemble.

Acknowledgments

This work is partly supported by the National Natural Science Foundation of China (Nos. 61170040, 61170040, 71371063) and the Foundation of Hebei University (No. 3504020).

References

[1] D. M. J. Tax, One-class classification: concept learning in the absence of counter examples, PhD Thesis, Felft University of Technology, 2001.

[2] O. Boehm, D. R. Hardoon, L. M. Manevitz, Classifying cognitive states of brain activity via one-class neural networks with feature selection by genetic algorithms, International Journal of Machine Learning and Cybernetics, 2(3) (2011) 125-134.

[3] L. V. Utkin, A framework for imprecise robust one-class classification models, International Journal of Machine Learning and Cybernetics, 5(3) (2014) 379-393.

[4] H. J. Shin, D. H. Eom, S. S.Kim, One-class support vector machine–an application in machine fault detection and classification, Computer & Industrial Engineering, 48(2) (2005) 395-408.

[5] G. Giacinto, R. Perdisci, M. Del Rio, F. Roli, Intrusion detection in computer networks by a modular ensemble of one-class classifiers, Information Fusion, 9(1) (2008) 69-82.

[6] J. Mourão-Miranda, D. R. Hardoon, T. Hahn, A. F. Marquand, S. C. R. Williams, J. Shawe-Taylor, M. Brammer, Patient classification as an outlier detection problem: an application of the one-class support vector machine, NeuroImage, 58(3) (2011) 793-804.

[7] K. Kennedy, B. M. Namee, S. J. Delany, Credit scoring: solving the low default portfolio problem using one-class classification, in: Proceedings of the 20th Irish Conference on Artificial Intelligence and Cognitive Science, 2009, pp. 168-177.

[8] C. He, M. Girolami, G. Ross, Employing optimized combinations of one-class classification for automated currency validation, Pattern Recognition, 37(6) (2004) 1085-1096.

[9] S. Cho, C. Han, D. Han, H. Kim, Web based keystroke dynamics identify verification using neural networks, Journal of Organizational Computing and Electronic Commerce, 10(4) (1997) 295-307.

[10] D. M. J. Tax, R. P. W. Duin, Support vector data description, Machine Learning, 54(1) (2004) 45-66.

[11] B. Schölkopf, The kernel trick for distances, in: Advances in Neural Information Processing Systems, vol. 13, 2001, pp. 301-307.

[12] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson, Estimating the support of a high-dimensional distribution, Neural Computation, 13 (2001) 1443-1471.

[13] D. M. J. Tax, R. P. W. Duin, Combining one-class classifiers, in: Proceedings of the 2nd International Workshop on Multiple Classifier Systems, 2001, pp. 299-308.

[14] S. Seguí, L. Igual, J. Vitrià, Weighted bagging for graph based one-class classifiers, in: N. El Gayar, J. Kittler, F. Roli (eds.), MCS 2010, LNCS, vol. 5997, 2010, pp. 1-10.

[15] J. Zhang, J. Lu, G. Q. Zhang, Combining one class classification models for avian influenza outbreaks, in: The 2011 IEEE Symposium on Computational Intelligence in Multicriteria Decision-Making, 2011, pp. 190-196.

[16] F. Hamdi, Y. Bennani, Learning random subspace novelty detection filters, in: The 2011 International Joint Conference on Neural Networks, 2011, pp. 2273-2280.

[17] T. Wilk, M. Wozniak, Soft computing methods applied to combination of one-class classifiers, Neurocomputing, 75 (2012) 185-193.

[18] B. Krawczyk, M. Woźniak, B. Cyganek, Clustering-based ensembles for one-class classification, Information Sciences, 264 (2014) 182-195.

[19] J. Liu, J. Song, Q. Miao, Y. Cao, FENOC: an ensemble one-class learning framework for malware detection, in: The 2013 Ninth International Conference on Computational Intelligence and Security, 2013, pp. 523-527.

[20] P. Casale, O. Pujol, P. Radeva, Approximate polytope ensemble for one-class classification, Pattern Recognition, 47 (2014) 854-864.

[21] T. G. Dietterich, An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization, Machine Learning, 40 (2000) 139-157.

[22] L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy, Machine Learning, 51 (2003) 181-207.

[23] G. Brown, L. I. Kuncheva, "Good" and "bad" diversity in majority vote ensembles, in: The 9th Inernational Workshop on Multiple Classifier Systems, 2010, pp. 124-133.

[24] P. Sidhu, M. P. S. Bhatia, An online ensembles approach for handling concept drift in data streams: diversified online ensembles detection. International Journal of Machine Learning and Cybernetics, 6(6) (2015) 883-909.

[25] P. Sidhu, M. P. S. Bhatia, A novel online ensemble approach to handle concept drifting data streams: diversified dynamic weighted majority, International Journal of Machine Learning and Cybernetics, 2015, DOI: 10.1007/s13042-015-0333-x.

[26] L. I. Kuncheva, Combining Pattern Classifiers. Methods and Algorithms, Wiley, New York (2004).

[27] L. I. Kuncheva, C. J. Whitaker, Measures of diversity in classifier ensembles, Machine Learning, 51 (2003) 181-207.

[28] B. Krawczyk, Diversity in ensembles for one-class classification, in : Current developments in databases and information systems, ser. Advances in Intelligent and Soft Computing, 2012, pp. 119-129.

[29] B. Krawczyk, M. Woźniak, Diversity measures for one-class classifier ensembles, Neurocomputing, 126 (2014) 36-44.

[30] B. Krawczyk, M. Woźniak, Accuracy and diversity in classifier selection for one-class classification ensembles, in: The 2013 IEEE Symposium on Computational Intelligence and Ensemble Learning, 2013, pp. 46-51.

[31] E. Menahem, L. Rokach, Y. Elovici, Combining one-class classifiers via meta learning, in: The 22nd ACM International Conference on Information and Knowledge Management, 2013, pp. 2435-2440.

[32] H. C. Kim, S. Pang, H. M. Je, D. Kim, S. Y. Bang, Constructing support vector machine ensemble, Pattern Recognition, 36 (2003) 2757-2767.

[33] H. H. Aghdam, E. J. Heravi, D. Puig, A new one class classifier based on ensemble of binary classifiers, in: The 16th International Conference on Computer Analysis of Images and Patterns, 2015, Part II: 242-253.

[34] I. Czarnowski, P. Jedrzejowicz, Ensemble online classifier based on the one-class base classifiers for mining data streams, Cybernetics and Systems, 46(1-2) (2015) 51-68.

[35] Z. H. Zhou, J. X. Wu, W. Tang, Ensembling neural networks: many could be better than all, Artificial Intelligence, 137(1-2) (2002) 239-263.

[36] N. Li, Z. Zhou, Selective ensemble under regularization framework, in: Lecture Notes in Computer Science, vol. 5519, 2009, pp. 293-303.

[37] L. Zhang, W. Zhou, Sparse ensembles using weighted combination methods based on linear programming, Pattern Recognition, 44(1) (2011) 97-106.

[38] Y. T. Yan, Y. P. Zhang, Y. W. Zhang, X. Q. Du, A selective neural network ensemble classification for incomplete data, International Journal of Machine Learning and Cybernetics, 2016, DOI: 10.1007/s13042-016-0524-0.

[39] X. Z. Wang, R. A. R. Ashfaq, A. M. Fu, Fuzziness based sample categorization for classifier performance improvement, Journal of Intelligent & Fuzzy Systems, 29(3) (2015) 1185-1196.

[40] X. Z. Wang, Learning from big data with uncertainty-editorial, Journal of Intelligent & Fuzzy Systems, 28(5) (2015) 2329-2330.

[41] Z. H. You, Y. K. Lei, L. Zhu, J. F. Xia, B. Wang, Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis, BMC Bioinformatics 14(Suppl 8) (2013): S10.

[42] Z. H. You, J. Z. Yu, L. Zhu, S. Li, Z. K. Wen. A mapreduce based parallel SVM for large-scale predicting proteinCprotein interactions, Neurocomputing, 145 (2014) 37-43.

[43] B. Krawczyk, M. Woźniak, Optimization algorithms for one-class classification ensemble pruning, in: The 6th Asian Conference on Intelligent Information and Database Systems, 2014, Part II: 127-136.

[44] B. Krawczyk, M. Woźniak, One-class classification ensemble with dynamic classifier selection, in: The 11th International Symposium on Neural Networks, 2014, pp. 542-549.

[45] B. Krawczyk, One-class classifier ensemble pruning and weighting with firefly algorithm, Neurocomputing, 150(B) (2015) 490-500.

[46] B. Krawczyk, Forming ensembles of soft one-class classifiers with weighted bagging, New Generation Computing, 33(4) (2015) 449-466.

[47] E. Parhizkar, M. Abadi, BeeOWA: A novel approach based on ABC algorithm and induced OWA operators for constructing one-class classifier ensembles, Neurocomputing, 166 (2015) 367-381.

[48] V. N. Vapnik, Statistical Learning Theory, Wiley, New York (1998).

[49] R. Jenssen, T. Eltoft, D. Erdogmus, J. C. Principe, Some equivalences between kernel methods and information theoretic methods, Journal of VLSI Signal Processing Systems, 49(1-2) (2006) 49-65.

[50] J. C. Principe, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives, Springer (2010).

[51] K. Torkkola, Feature extraction by non-parametric mutual information maximization, Journal of Machine Learning Research, 3 (2003) 1415-1438.

[52] X. T. Yuan, B. G. Hu, Robust feature extraction via information theoretic learning, in: Proceedings of the 26th Annual International Conference on Machine Learning, 2009, pp. 1193-1200.

[53] R. Rockfellar, Convex Analysis, Princeton University, Princeton, NJ, 1970.

[54] S. Yang, H. Zha, S. Zhou, B. Hu, Variational graph embedding for globally and locally consistent feature extraction, in: Proceedings of the Europe Conference on Machine Learning, 2009, pp. 538-553.

[55] W. Liu, P. P. Pokharel, J. C. Principe, Correntropy: properties and applications in non-Gaussian signal processing, IEEE Transactions on Signal Processing, 55(11) (2007) 5286-5297.

[56] K. H. Jeong, W. Liu, S. Han, E. Hasanbelliu, J. C. Principe, The correntropy MACE filter, Pattern Recognition, 42(5) (2009) 871-885.

[57] S. M. Guo, L. C. Chen, J. S. H. Tsai, A boundary method for outlier detection based on support vector domain description, Pattern Recognition, 42(1) (2009) 77-83.

[58] M. Wu, 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, 31(11) (2009) 2088-2092.

[59] L. Breiman, Bagging predictors, Machine Learing, 24(2) (1996) 123-140.

[60] Y. Freund, R. E. Schapire, A short introduction to boosting, Journal of Japanese Society for Artificial Intelligence, 14(5) (1999) 771-780.

[61] V. Cheplygina, D. M. J. Tax, Pruned random subspace method for one-class classifiers, in: The 10th International Workshop on Multiple Classifier Systems, 2011, pp. 96-105.

[62] A. Frank, A. Asuncion, UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences, Irvine, CA, 2010.

[63] H. J. Xing, B. G. Hu, An adaptive fuzzy c-means clustering-based mixtures of experts model for unlabeled data classification, Neurocomputing, 71 (2008) 1008-1021.

[64] B. W. Silverman, Density Estimation for Statistics and Data Analysis, Chapman and Hall, London, U.K., 1986.

[65] B. Krawczyk, P. Filipczuk, Cytological image analysis with firefly nuclei detection and hybrid one-class classficiation decomposition, Engineering Applications of Artificial Intelligence, 31 (2014) 126-135.

[66] S. Kang, S. Cho, P. Kang, Multi-class classification via heterogeneous ensemble of one-class classifiers, Engineering Applications of Artificial Intelligence, 43 (2015) 35-43.

[67] B. Krawczyk, M. Woźniak, F. Herrera, On the usefulness of one-class classifier ensembles for decomposition of multi-class problems, Pattern Recognition, 48(12) (2015) 3969-3982.