Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A SMALL SPHERE AND PARAMETRIC VOLUME FOR SUPPORT VECTOR DOMAIN DESCRIPTION

[1]MOHAMED EL BOUJNOUNI, [2]MOHAMED JEDRA, [3]NOUREDDINE ZAHID
[1]Doctoral Student, Conception & Systems Laboratory, FSR, Morocco
[2]Prof, Conception & Systems Laboratory, FSR, Morocco
[3]Prof, Conception & Systems Laboratory, FSR, Morocco
E-mail: [1]med_elbouj@yahoo.fr , [2] jedra@fsr.ac.ma , [3]zahid@fsr.ac.ma

ABSTRACT

Support Vector Domain Description (SVDD) is inspired by the Support Vector Classifier. It obtains a sphere shaped decision boundary with minimal volume around a dataset. This data description can be used for novelty or outlier detection. Our approach is always to minimize the volume of the sphere describing the dataset, but following the value of a parameter, which controls its volume and plays a compromise between the outlier's acceptance and the target's rejection. Simulation results on seven benchmark datasets have successfully validated the effectiveness of the proposed method.

Keywords: *Parametric Volume, Small Hypersphere, Compromise.*

1 INTRODUCTION

In domain description the task is not to distinguish between classes of objects like in clustering problems, or to produce a desired outcome for each input object like in regression problems, but to give a description of a set of objects. This description should be able to distinguish between the classes of objects represented by the training set, and all other possible objects in the object space [6 9]. Recently, a Support Vector Domain Description (SVDD) (also called One-class Classification), inspired by support vector machine was invented by Tax and Duin [9 10 11]. In a SVDD the compact description of target data is given as a hypersphere with minimal volume containing most of normal data, rejecting most of negative data. It has the possibility of transforming the data to new feature spaces without much extra computational cost using kernel functions, [4 7]. SVDD uses support vectors to describe the boundary of target class as Support Vector Machine SVM does [2 13]. Support vectors are found by solving convex quadratic programming (QP) problem [3 5], when all support vectors are calculated, the hyperspheres associated with each class is determined by the couple (center, radius). In the decision phase a sample is classified into class i only when the value of the i^{th} decision function is positive. This automatic architecture does not permit to control the volume of the hyperspheres. Our aim is to integrate a reel and positive parameter called p, in standard SVDD. p is a compromise between the outlier's acceptance and the target's rejection, so a good choice of p can improve the classification. It resolve also a problem linked to the classification of the set of Support Vectors (SV's), because following the KKT (Karush-Kuhn-Tucker) conditions, the decision function gives a null value (not signed) for (SV's), but in some cases it can contains both of (target and outlier) , consequently, the decision function can not distinguish between them .

The rest of this paper is organized as follows: in section 2, we present the Support Vector Domain Description algorithm. Our approach is explained in section 3. Experimental results are provided in section 4 and finally we conclude the paper in the last section.

2 SUPPORT VECTOR DOMAIN DESCRIPTION

2.1 Definition and Formulation

The normal data description model [1 11 15 17] gives a closed boundary around the data: a sphere characterized by center a and radius R > 0. It minimizes the volume of the sphere by minimizing R^2, and demand that the sphere contains all training objects.

Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Let $\{x_i\} \in \chi$ be a dataset of N points, with, $x_i \in R^d$ the data space, we look for the smallest enclosing sphere of radius R which is described by the following constraints:

$$\|x_j - a\|^2 \le R^2 \qquad \forall j \tag{1}$$

Where $\|.\|$ is the Euclidean norm. Soft constraints are incorporated by adding slack real and positive variable ε_j :

$$\|x_j - a\|^2 \le R^2 + \varepsilon_j \quad \forall j \tag{2}$$

To solve this problem we introduce the Lagrangian :

$$L = R^2 - \sum_j (R^2 + \varepsilon_j - \|x_j - a\|^2)\alpha_j - \sum_j \varepsilon_j \mu_j + C\sum_j \varepsilon_j \tag{3}$$

Where $\alpha_j \ge 0$ and $\mu_j \ge 0$ are Lagrange multipliers, C is a constant, and $C\sum_j \varepsilon_j$ is a penalty term. Setting the partial derivatives of L with respect to R, a, ε_i to zero gives the following constraints:

$$\partial_R L = 0 \Rightarrow \sum_i \alpha_i = 1 \tag{4}$$

$$\partial_a L = 0 \Rightarrow a = \sum_i \alpha_i x_i \tag{5}$$

$$\partial \varepsilon_i L = 0 \Rightarrow \alpha_i = C - \mu_i \tag{6}$$

The solution of the primal problem can be obtained by solving its dual problem [11]:

Max:

$$W = \sum_j x_j^2 \alpha_j - \sum_{i,j} \alpha_i \alpha_j x_i x_j$$
$$Subject\ to: \tag{7}$$
$$0 \le \alpha_j \le C \quad \forall j \quad and \quad \sum_j \alpha_j = 1$$

When negative examples (objects which should be rejected) are available, they can be incorporated in the training to improve the description. In contrast with the training (target) examples which should be within the sphere, the negative examples should be outside it. In the following, the target objects are enumerated by indices i, j and the negative examples by *l*, m. Again we allow for errors in both the target and the outliers set and introduce slack real positive variables ε_i and ε_l [11]:

$$L(R, \varepsilon_i, \varepsilon_l) = R^2 + C1\sum_i \varepsilon_i + C2\sum_l \varepsilon_l \tag{8}$$

With the constraints:

$$\|x_i - a\|^2 \le R^2 + \varepsilon_i \quad \|x_l - a\|^2 \ge R^2 - \varepsilon_l \quad \varepsilon_i \ge 0 \quad \varepsilon_l \ge 0 \quad \forall i,l \tag{9}$$

Where C1, C2 are constants real positives, and $C1\sum_i \varepsilon_i$, $C2\sum_l \varepsilon_l$ are penalty terms, These constraints are incorporated in (eq 8) and the Lagrange multipliers α_i, α_l ,γ_i, γ_l are introduced as follows:

$$L(R,a,\varepsilon_i,\varepsilon_l,\alpha_i,\alpha_l,\gamma_i,\gamma_l) = R^2 + C1\sum_i \varepsilon_i + C2\sum_l \varepsilon_l - \sum_i \gamma_i \varepsilon_i - \sum_l \gamma_l \varepsilon_l$$
$$- \sum_i \alpha_i [R^2 + \varepsilon_i - \|x_i - a\|^2] - \sum_l \alpha_l [\|x_l - a\|^2 - R^2 + \varepsilon_l] \tag{10}$$

With $\alpha_i \ge 0, \alpha_l \ge 0, \gamma_i \ge 0, \gamma_l \ge 0$ are Lagrange multipliers. Setting the partial derivatives of L with respect to R, a, ε_i and ε_l to zero gives the following constraints:

$$\partial_R L = 0 \Rightarrow \sum_i \alpha_i - \sum_l \alpha_l = 1 \tag{11}$$

$$\partial_a L = 0 \Rightarrow a = \sum_i \alpha_i x_i - \sum_l \alpha_l x_l \tag{12}$$

$$\partial_{\varepsilon_i} L = 0 \text{ and } \partial_{\varepsilon_l} L = 0 \Rightarrow \alpha_i = C1 - \gamma_i \quad \alpha_l = C2 - \gamma_l \quad \forall i,l \tag{13}$$

When (eqs 11,12,13) are substituted into (eq 10) we obtain :

Max:

$$W = \sum_i \alpha_i (x_i x_i) - \sum_l \alpha_l (x_l x_l) - \sum_{i,j} \alpha_i \alpha_j (x_i x_j) + 2\sum_{l,j} \alpha_l \alpha_j (x_l x_j) - \sum_{l,m} \alpha_l \alpha_m (x_l x_m)$$
$$Subject\ to: 0 \le \alpha_i \le C1 \quad and \quad 0 \le \alpha_l \le C2 \quad \forall i,l$$
$$\sum_i \alpha_i - \sum_l \alpha_l = 1 \tag{14}$$

The formulations of SVDD can be extended to obtain a more flexible description. Data is mapped nonlinearly into a higher dimensional space where a hyperspherical description can be found. The mapping is performed implicitly, replacing all of the inner products by a kernel function K (xi, xj) [1 8 11 16]. Table 1 describes some commonly used kernel functions:

Table 1 Some Commonly Used Kernel Functions

Kernel's types	K(x, x_j)
Polynomial	$(1 + x^T.x_i)^p$
Radial-basis function	$\exp(-\frac{\|x - x_i\|^2}{2\sigma^2})$
Two-layer perception	$Tanh(s_0 x^T.x_i + s_1)$

Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

2.2 Decision Function

To classify a test point z, we just investigate whether it is inside the hypersphere (a_k,R_k) constructed during the training and associated to the class k [8 9 10 11] .Namely the decision function is calculated as (eq 15), if its value is positive for the k[th] class and negative for the others we conclude that z belong to the class k.

$$f(z) = \mathrm{sgn}(R_k^2 - \|z - a_k\|^2) \quad (15)$$

This function can be calculated as follows:
In the normal data description case we obtain:

$$\|z - a_k\|^2 = z.z - 2\sum_i \alpha_{ki} x_i z + \sum_{i,j} \alpha_{ki}\alpha_{kj} x_i x_j \quad (16)$$

$$R_k^2 = x.x - 2\sum_i \alpha_{ki} x_i x + \sum_{i,j} \alpha_{ki}\alpha_{kj} x_i x_j \quad (17)$$

With α_{kj} is the j[th] Lagrangian multiplier corresponding to the k[th] cluster, and $x \in SV$ the set of Support Vectors having $0 \prec \alpha_i \prec C$.

In the SVDD with negative examples case we obtain:

$$\|z - a_k\|^2 = z.z - 2\left[\sum_i \alpha_{ki} x_i z - \sum_l \alpha_{kl} x_l z\right] + \sum_{i,j} \alpha_{ki}\alpha_{kj} x_i x_j + \sum_{l,m} \alpha_{kl}\alpha_{km} x_l x_m - 2\sum_{i,l} \alpha_{ki}\alpha_{kl} x_i x_l \quad (18)$$

$$R_k^2 = x.x - 2\left[\sum_i \alpha_{ki} x_i x - \sum_l \alpha_{kl} x_l x\right] + \sum_{i,j} \alpha_{ki}\alpha_{kj} x_i x_j + \sum_{l,m} \alpha_{kl}\alpha_{km} x_l x_m - 2\sum_{i,l} \alpha_{ki}\alpha_{kl} x_i x_l \quad (19)$$

For any $x \in SV$ the set of support vectors having $0 \prec \alpha_i \prec C1$ (with x is a target object) or $0 \prec \alpha_l \prec C2$ (with x is negative object).

3 OUR APPROACH : SMALL SPHERE AND PARAMETRIC VOLUME

3.1 Definition and Formulation

To integrate the parameter p in SVDD we have to reformulate the optimization problem by introducing p in the constraints. To simplify the description, we use the same mathematical calculation like SVM [13 14] does. Suppose we are given a dataset of N points (x_1,y_1)… (x_n,y_n), with $x_i \in R^d$, the data space, and y_i =1 for the positive examples, -1 for the negative ones. We look for the smallest sphere of radius R that encloses only the positive samples, which is described by the following constraints:

$$\|x_j - a\|^2 \le R^2 - p.y_j + \varepsilon_j \ \forall j \quad \text{with} \quad y_j = +1 \quad (20)$$

$$\|x_j - a\|^2 \ge R^2 - p.y_j - \varepsilon_j \ \forall j \quad \text{with} \quad y_j = -1 \quad (21)$$

When p is equal to 0 the results are the same as the conventional SVDD, because the two constraints above will be the same as (eq 9)

To solve this problem we introduce the Lagrangian:

$$L = R^2 + C\sum_j \varepsilon_j - \sum_j \alpha_j.y_j.\left(R^2 - \|x_j - a\|^2 - p.y_j\right) - \sum_j \alpha_j\varepsilon_j - \sum_j \varepsilon_j\mu_j \quad (22)$$

Setting the partial derivatives of L with respect to R, a, ε_j to zero gives the following constraints:

$$\partial_R L = 0 \Rightarrow \sum_i \alpha_i.y_i = 1 \quad (23)$$

$$\partial_a L = 0 \Rightarrow a = \sum_i \alpha_i x_i y_i \quad (24)$$

$$\partial_{\varepsilon_i} L = 0 \ \Rightarrow \ \alpha_i = C - \mu_i \quad \forall i \quad (25)$$

Hence, the dual optimization problem becomes Max:

$$w = -\sum_{i,j} \alpha_i\alpha_j y_i y_j x_i x_j + \sum_j \alpha_j (y_j x_j x_j + p) \quad \text{subject to} \quad 0 \le \alpha_i \le C \quad \text{and} \quad \sum_i \alpha_i y_i = 1 \quad (26)$$

3.2 Decision function

To classify a sample z, we use the same decision function as SVDD (eq 15), but with a value of radius depending on the parameter p as follows:

$$R_k^2 = \|x_l - a_k\|^2 + y_l.p \ \forall j \quad (27)$$

$$R_k^2 = x_l.x_l - 2\sum_j x_j x_i y_j \alpha_j + \sum_{i,j} \alpha_i\alpha_j y_i y_j x_i x_j + y_l.p \quad (28)$$

$x_l \in SV$, the set of Support Vectors having $0 \prec \alpha_l \prec C$, R_k and a_k are respectively the radius and the center of the k[th] class.

To illustrate the role and the values of the parameter p, we run Small Sphere and Parametric Volume for SVDD on two different datasets (see Fig.1) , using polynomial and Gaussian kernels, with varying degrees and widths respectively, C is equal to 1000. In each test we fix the values of sigma (width) or d, and we increase gradually the value of p, beginning by 0; we remind that zero corresponds to the standard SVDD. The layers with the gray scale from white to black, represents the increase of p. The values of p corresponding to

Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

each layer are indicated in Table 2. We observe that: All layers follow the form of the distribution of the points of each class. As far as sigma or d increases the surface of layers grows, For the RBF kernel a small modification of p (see Table 2), produces the appearance of new layers, contrary to the polynomial kernel. We remark also that when p takes large values the layers of each class can overlap.

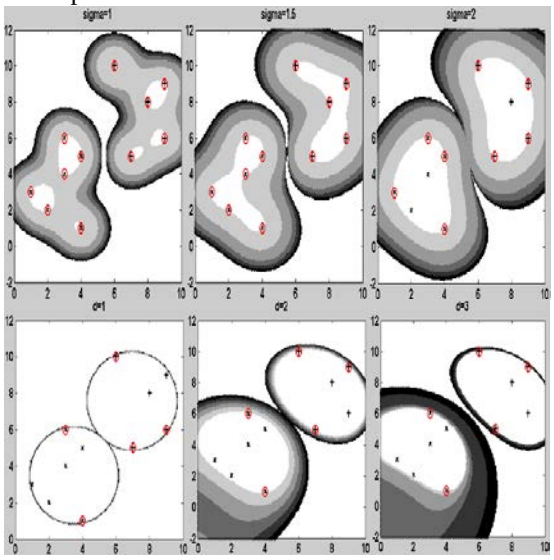


Fig. 1 A Small Sphere And Parametric Volume For SVDD Trained On An Artificial Dataset. Support Vectors Are Indicated By The Reds Circles (Just When P Is 0). The Layers With The Gray Scale From White To Black Correspond To The Increase Of P.

Table 2 Numerical Values Of The Parameter P, Corresponding To The Polynomial And Gaussian Kernels Used In The Experiment Above

Kernel	Gaussian			Polynomial		
	σ=1	σ=1.5	σ=2	d=1	d=2	d=3
The values of p	0	0	0	0	0	0
	0.02	0.02	0.02	0.1	150	1000
	0.04	0.04	0.04	0.2	300	5000
	0.06	0.06	0.06	0.3	450	10000
	0.08	0.08	0.08	0.4	600	50000
	1	1	1	0.5	750	100000

4 EXPERIMENTAL RESULTS

4.1 Datasets and Experimental Setting

Seven datasets were used to test the Small Sphere and Parametric Volume for SVDD. The datasets describe various characteristics from studying monk-1, monk-2, monk-3, iris flowers, wine, glass, and ecoli; all of these datasets are taken from [12], Further details of these datasets are provided in Table 3.

Table 3 Description Of The Datasets Used In The Experiment, Training Samples And Testing Samples List The Rate Of Data Used Or Directly The Files Containing Data

dataset	Number of data	subset		Number of class	Feature
		Training set	Testing set		
monks	432	monks-1.train	monks-1.test	2	6
	432	monks-2.train	monks-2.test	2	6
	432	monks-3.train	monks-3.test	2	6
iris	150	80% of samples/each class	The remaining samples/each class	3	4
wine	178	80% of samples/each class	The remaining samples/each class	3	13
glass	214	80% of samples/each class	The remaining samples/each class	6	9
e coli	336	80% of samples/each class	The remaining samples/each class	8	7

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Firstly, the three problems defined for monk's dataset were used in the experiment; monks-1 is in standard disjunctive normal form and is supposed to be easily learnable by most of the algorithms and decision trees. Conversely, monk's-2 is similar to parity problems. It combines different attributes in a way that makes it complicated to describe using the given attributes only; monks-3 serves to evaluate the algorithms under the presence of noise.

Secondly, the iris dataset consists of three classes, each of which has 50 samples. While one cluster is easily separable, it is difficult to achieve separation between the other two clusters. Data points correspond to the plants and attributes correspond to sepal and petal measurements.

Thirdly, the wine dataset is the results of a chemical analysis of wines grown in the same region but derived from three different cultivars. The analysis determined the quantities of constituents found in each of the three types of wines.

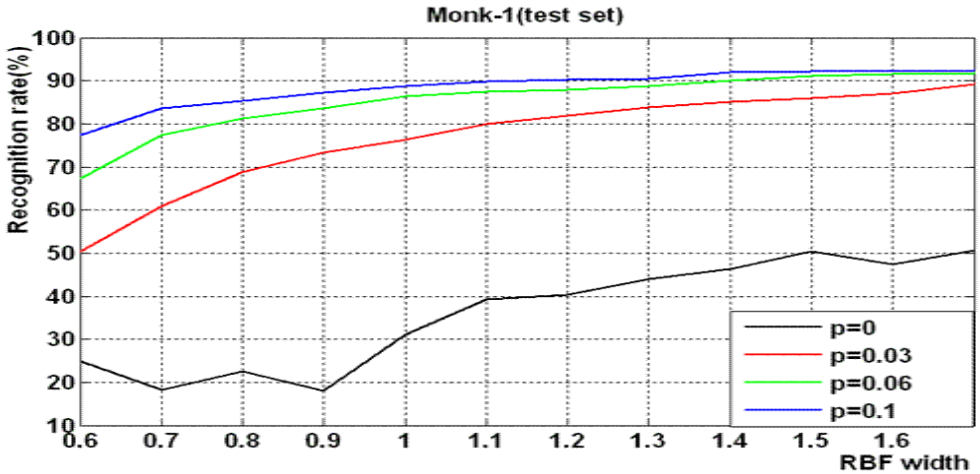
Fourthly, the glass dataset is the study of classification of types of glass was motivated by criminological investigation data points correspond to the type of glass and attributes correspond to their oxide content (i.e. Na, Fe, K, etc).

Fifthly, the e coli dataset is related to protein localization sites, it contains 336 patterns divided into 8 classes: class 1 (cytoplasm), class 2 (inner membrane without signal sequence), class 3 (perisplasm), class 4 (inner membrane, uncleavable signal sequence), class 5 (outer membrane), class 6 (outer membrane lipoprotein), class 7 (inner membrane lipoprotein) and class 8 (inner membrane cleavable signal sequence).

For monk's problem we use the files monks-(1, 2, 3).train, as training set and their corresponding files monks-(1, 2, 3).test, as testing set. For iris, wine, glass and ecoli datasets, we randomly split each one into 20 subsets, each subset contains training and testing sets, with the scheme described in Table 3.Training and test sets do not intersect.

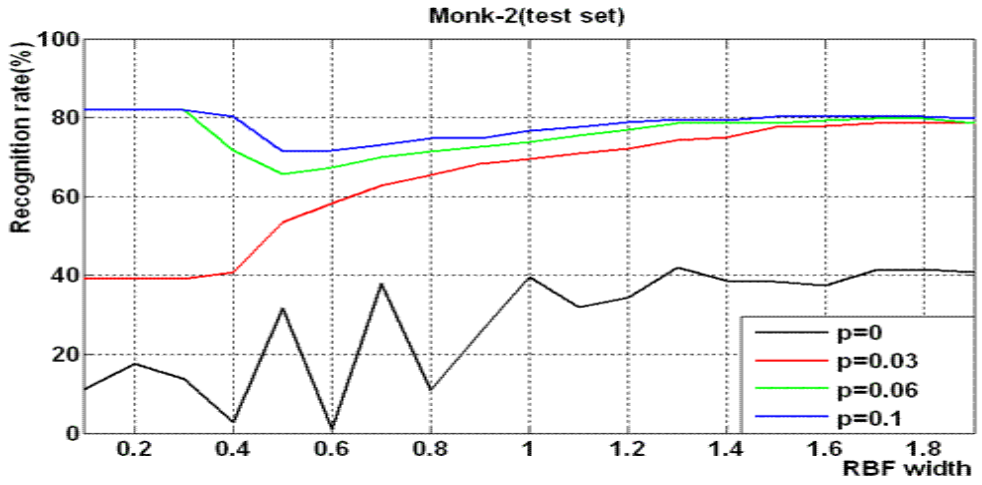
4.2 Numerical Results

For iris, wine, glass, and e coli: To test a dataset, we select the values of p and RBF width, we fix C=1000, then for each subset of this dataset the algorithm Small Sphere and Parametric Volume for SVDD will be trained by the training dataset and then, tested by the training and the testing dataset. After terminating the 20 experiments, we calculate the average and the standard deviation.
For monks (1 2 3) datasets: We select the values of p and RBF width, we fix C=1000, the algorithm Small Sphere and Parametric Volume for SVDD will be trained by monks-x.train and then, tested by monks-x.train, and monks-x.test, where x takes the values {1,2,3}.

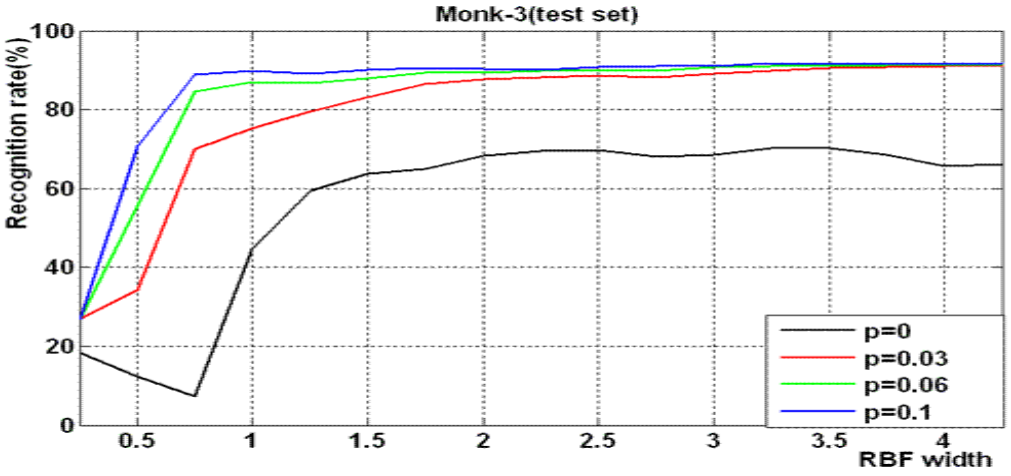


(a)

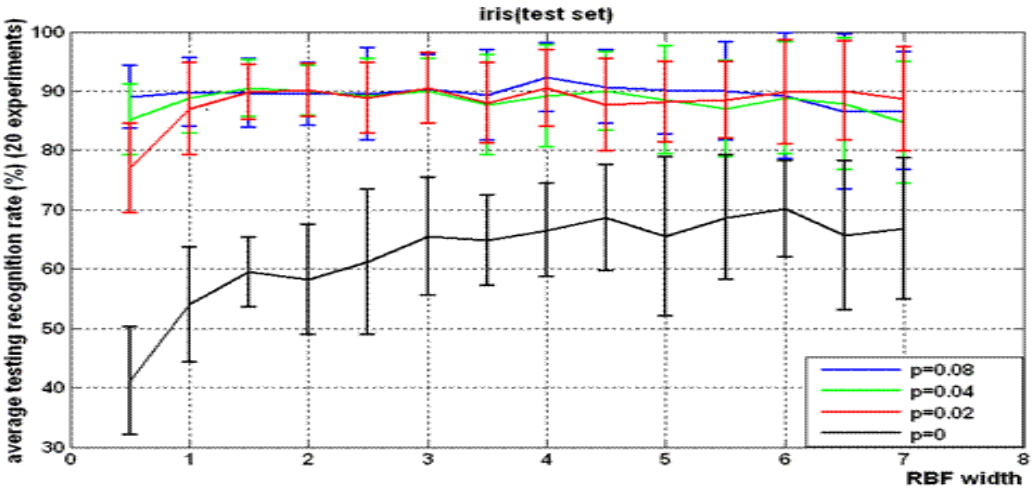
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(b)



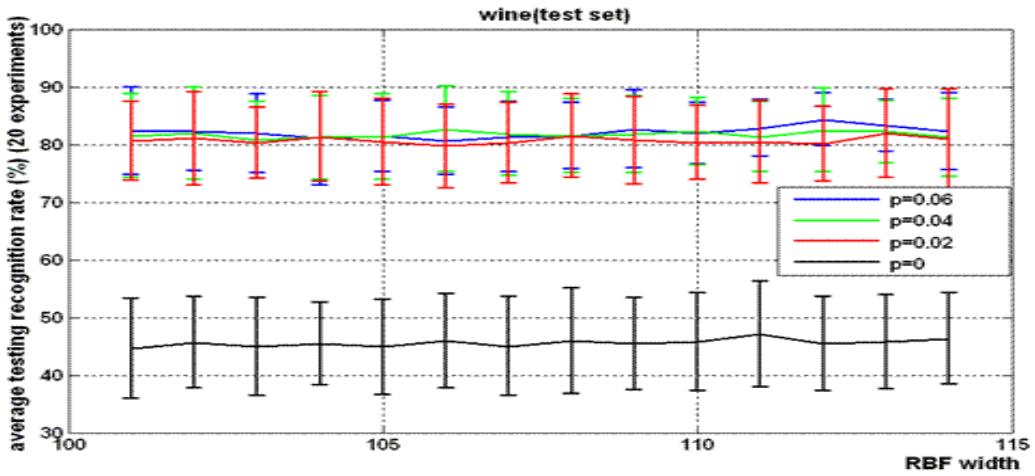
(c)



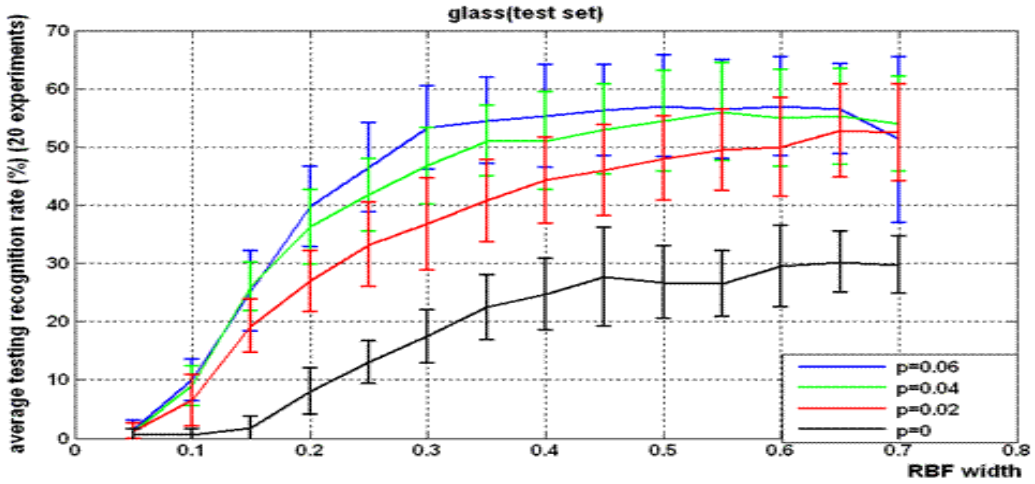
(d)

Journal of Theoretical and Applied Information Technology
15th December 2012. Vol. 46 No.1

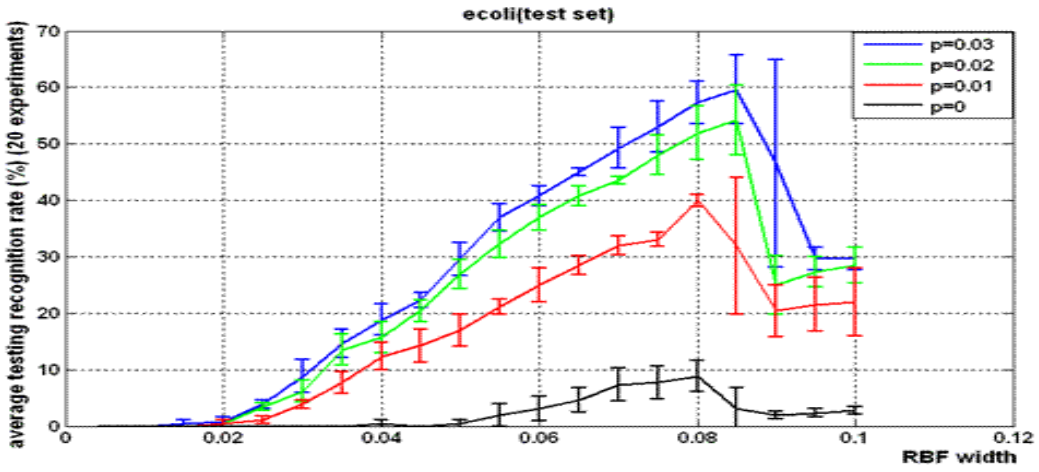
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195



(e)



(f)



(g)

Fig. 2 Recognition Rate And Standard Deviation (Indicated By Symmetric Error Bars) For Test Set, Using (A) Monk-1, (B) Monk-2, (C) Monk-3, (D) Iris, (E) Wine, (F) Glass, And (G) Ecoli Dataset, Against RBF Width

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

At first we inform that: for the seven experiments above, when using the training dataset as test, the recognition rate achieves 100%, whatever the value of the parameter p.

Fig 2 shows the recognition rate of the testing datasets against the RBF width (σ), using different values of p. In all experiments we remark that: when p takes not null values, the recognition rate increases whatever the value of σ. For a fixed σ the recognition rate grows as the value of p is raised, and as far as sigma increases, the recognition rate grows until a maximum value. This result shows that, the use of this parameter p achieves better performance than standard SVDD, and gives a good separability between the different classes.

5 CONCLUSION

A Small Sphere and Parametric Volume for SVDD was developed to control the volume of the hypersphere characterizing each class, by integrating a real and positive parameter p in SVDD , that plays a compromise between the outlier's acceptance and the target's rejection, by consequence a good choice of the value of p can achieve better performance on improving the recognition rate , it allows also to classify the set of support vector, when it contains simultaneously positive and negative samples. The proposed method has been implemented, analyzed, and tested on seven benchmark datasets. The results obtained are very encouraging and, it provides better recognition rate than the SVDD without using of the parameter p.

REFRENCES:

[1] Asa Ben-Hur, David Horn, Hava T.Siegelmann, Vladimir Vapnik ,"Support vector clustering", *Journal of Machine Learning Research*. Vol. 2, No. 12, 2001, pp. 125-137.

[2] Cortes, C., and V. Vapnik,"Support-vector networks". *Machine Learning*, Vol. 20, No. 3, 1995, pp. 273–297.

[3] Kim, P.J, H.J Chang, D.S Song, and J.Y Choi. "Fast Support Vector Data Description Using K-Means". *LNCS 4493*, 2007, pp 506–514.

[4] Lee, K., D.W Kim, D. Lee, and K. H Lee. "Improving support vector data description using local density degree". *Pattern Recognition*, Vol. 38, No. 10, 2005, pp. 1768 – 1771.

[5] Platt, J.C . "Fast Training of Support Vector Machines using Sequential Minimal Optimization". *Advances in Kernel Methods – Support Vector Learning. MIT Press*, Cambridge, 1999. pp. 41–65.

[6] Ritter, G., and M. Gallegos, "Outliers in statistical pattern recognition and an application to automatic chromosome classification". *Pattern Recognition Letters*, Vol. 18, 1997, pp 525–539.

[7] Schölkopf, B., and A.J. Smola." Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond". *Cambridge Mass: MIT Press, London*, 2002.

[8] Tavakkoli, A., M. Nicolescu, G. Bebis, and M. Nicolescu . "A support vector data description approach for background modeling in videos with quasi-stationary backgrounds". *International journal on artificial intelligence tools*,Vol. 17, No. 4, 2008, pp. 635–658

[9] Tax, D., and R. Duin. "Data Domain Description Using Support Vectors". *Proceedings-European Symposium on Artificial Neural Networks Bruges Belgium*, 1999, pp. 251-256.

[10] Tax, D., and R. Duin. "Support vector domain description". *Pattern Recognition Letters*, Vol. 20, No. 11–13, 1999, pp. 1191–1199.

[11] Tax, D., and R. Duin, "Support Vector Data Description", *Machine Learning*. Vol. 54, 2004, pp. 45–66.

[12] UCI repository of machine learning databases. Http://archive.ics.uci.edu/ml/.

[13] Vapnik, V. "The Nature of Statistical Learning Theory". *Springer-Verlag*, New York, 1995.

[14] Vapnik, V. "Statistical Learning Theory". *Wiley, New York*, 1998.

[15]Wu, M. and J.Ye. "A Small Sphere and Large Margin Approach for Novelty Detection Using Training Data with Outliers". IEEE transactions on pattern analysis and machine intelligence, Vol. 31, No.11, 2009, pp. 2088-2092

[16] K.-R. Mäller, S. Mika, G. Rätsch, K. Tsuda, B. Schölkopf, "An introduction to kernel-based learning algorithms", *IEEE Transactions on Neural Networks*, Vol. 12, No. 2, 2001, pp. 181- 201.

[17] Lee, J., Lee, D., "An improved cluster labeling method for support vector clustering". *IEEE Trans. Pattern Analysis Machine Intelligence* Vol. 27 , 2005, pp. 461–464