Minimum Enclosing and Maximum Excluding Machine for Pattern Description and Discrimination

Yi Liu and Yuan F. Zheng
Department of Electrical and Computer Engineering
The Ohio State University
Columbus, Ohio 43210
Email:{liuyi, zheng}@ece.osu.edu

Abstract

This work addresses the description problem of a target class in the presence of negative samples or outliers. Traditional Support Vector Machines (SVM) has strong discrimination capability to distinguish the target classes but does not reject the uncharacteristic patterns well. The one-class SVM, on the other hand, provides good representation for the classes of interest but overlooks the discrimination issue between them. This paper presents a new one-class classifier named minimum enclosing and maximum excluding machine (MEMEM), which offers capabilities for both pattern description and discrimination. The properties of MEMEM are analyzed and the performance comparisons using synthetic and real data are presented.

Index Terms

Support Vector Machines (SVM), one-class SVM, description, discrimination, the minimum enclosing ball.

I. Introduction

This work addresses the description problem of a target class in the presence of negative samples or outliers. A typical application is face recognition and verification. The system needs to identify one or multiple clients (recognition) and perform reliable rejection of impostors (verification) as well. Available are some photos of the persons of interest and few, or most likely no, photos of potential impostors which evidently are far from representative for the non-target class. Similar applications include speaker identification, fingerprint verification, target recognition in SAR imagery and etc..

Most often those tasks are handled as a conventional classification problem. Aiming at distinguishing one class from others, the classification systems assume that every sample comes from one of the classes that have been fixed during the training process. As a result, difficulties arise when rejection output is required or some classes are severely undersampled, which unfortunately is the case for the problem under study. Even Support Vector Machines (SVM), a proven powerful tool for discrimination and has been popularly chosen to achieve the classification task for a wide range of applications, is known as not be able to reject uncharacteristic samples well and therefore suffers from false alarms [1].

A very simple scheme using output threshold has been extensively adopted to achieve the task of patten verification. However, as pointed out in [2], simply using a threshold does not adequately solve the problem. For this reason, the problem of class description or so-called one-class classification (OCC) has attracted attention of many researchers [3]–[7]. By modeling the *support* of the class where the data predominantly reside, OCC can recognize the new samples that resemble the training set and detect uncharacteristic samples, or outliers, to avoid the ungrounded classification. Capable of working even when the training samples are solely from the target class, OCC is especially valuable when it is difficult or expensive to collect the samples from the non-target class. By far, the well-known examples of OCC are studied in the context of SVM. The one-class SVM proposed by Tax and Duinin [4] is named support vector domain description (SVDD), which seeks the minimum hypersphere that encloses all the data in the target class. In this way, it finds the descriptive area that covers the data and excludes the superfluous space that results in false alarms. Two alternative approaches are developed in [5] and [6] which use hyperplanes to describe the class. One common limitation of the two methods, however, is the restriction for the samples to be unit norm.

In many systems, a two-step procedure is designed. First, the decision of acceptance or rejection is made by an OCC algorithm. Then the accepted sample will be further classified by conventional classifiers such as SVM into one of the predefined classes. The difficulty, however, is that the samples drawn from different classes may show great diversities and therefore do not necessarily cluster well. As a result, describing them as a whole by a single OCC is not able to deliver desired performance. A one-step approach is proposed in [8], which constructs M one-class SVM with one for each class. If all classifiers reject the input sample, the system rejects it; otherwise, it is assigned to the class with the highest confidence level of acceptance. Unlike the two-step procedure, each one-class classifier is now responsible for both recognition and rejection. However, because of the descriptive nature of OCC, the factor of the separation between classes is not considered in the formulation and therefore the discrimination issue is overlooked. In this work, we propose a novel one-class classifier in the context of SVM, which is named the minimum enclosing and maximum excluding machine (MEMEM). Similar to SVDD, MEMEM models the support of the target class by a hypersphere, but unlike SVDD it seeks the hypersphere that excludes the negative samples by a wide shell in addition to enclosing the target samples with small radius. By doing so, the discriminating ability of the classifier is enhanced while its descriptive ability is preserved.

The rest of the paper is organized as follows. Section II presents the formulation of MEMEM. Then experimental results are presented in section III which is followed by the conclusions and discussions in section IV.

II. MEMEM

For the problem of class description, an assumption is usually made that the separation boundary drawn by the classifier is closed [2]. SVDD, for example, imposes a hypersphere $B(a, R)$, which is characterized by the center a and radius R, as the separating surface around the samples. The spherical shape of the boundary, however, is a rigid model which fortunately can be made flexible by using kernel functions as by SVM. For its simplicity, we also models the boundary as a hypersphere.

Suppose we are given N training samples (x_i, y_i) with $x_i \in R^d$ and $y_i \in \{1, -1\}$, where the target class is defined as class 1. When the training samples are spherically separable, two concentric hyperspheres can be found such that all the samples from class 1 are enclosed by the inner hypersphere $B(a, R_1)$ and all the samples from class -1 are excluded by the outer hypersphere $B(a, R_2)$ as

$$
\begin{aligned}
||a - x_i||^2 &\leq R_1^2 \quad \text{for} \quad y_i = 1, \\
||a - x_i||^2 &\geq R_2^2 \quad \text{for} \quad y_i = -1. \qquad (1)
\end{aligned}
$$

Similar to SVDD, we want the inner hypersphere $B(a, R_1)$ to be as small as possible for good description of the target class. In the meantime, to separate the two classes we want the outer hypersphere $B(a, R_2)$, which pushes the negative samples away, to be as large as possible, and we call the one with the largest radius the maximum excluding sphere. Recall in SVM, the discrimination between two classes is achieved by maximizing the margin as shown in Fig. 1(a). Inspired by this idea, to obtain good discriminating ability for the hyperspheres we maximize the width or equivalently the area of the shell between $B(a, R_1)$ and $B(a, R_2)$ (the shaded area in Fig. 1(b)), which is proportional to the quantity $R_2^2 - R_1^2 \triangleq 2\Delta R^2$. Then the resulted boundary would be $B(a, R)$ where $R = \sqrt{(R_1^2 + R_2^2)/2}$ as shown as the dashed line in Fig. 1(b).

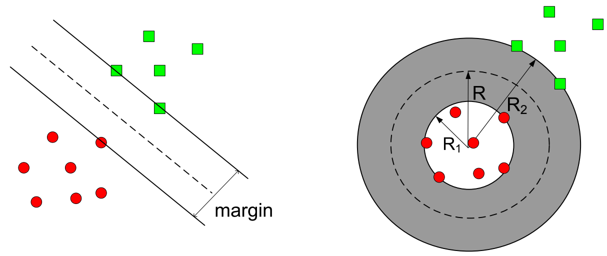
The simplest way to achieve the objective of both description (small R_1^2) and discrimination (large ΔR^2) is to minimize the quantity $R_1^2 - (R_2^2 - R_1^2) = R^2 - 3\Delta R^2$, or equivalently

$$\frac{1}{3} R^2 - \Delta R^2. \qquad (2)$$

With the motivations explained above, we now propose MEMEM as

$$
\begin{aligned}
\min_{a, R^2, \Delta R^2} \quad &: \quad \gamma R^2 - \Delta R^2 \\
\text{subject to} \quad &: \quad ||a - x_i||^2 \leq R^2 - \Delta R^2 \text{ for } y_i = 1, \\
&\quad\;\; ||a - x_i||^2 \geq R^2 + \Delta R^2 \text{ for } y_i = -1, \qquad (3)
\end{aligned}
$$

where a non-negative parameter γ replaces the constant $\frac{1}{3}$ in Eq. (2) to give users the flexibility to balance the importance of a small enclosing ball and a large shell. When the training samples are scarce, γ should be set close to 1. One extreme is $y_i = 1$ for $i = 1, \ldots, N$, which means no negative samples are available. In this case, γ should be equal to 1 and MEMEM only cares about the minimization of the enclosing hypersphere, which reduces it to SVDD. On the other hand, if the classes are well sampled such that the decision boundary can be supported



(a) (b)

Fig. 1. Discrimination between two classes. (a) SVM. (b) MEMEM.

from both sides, γ should be close to 0. By doing so, MEMEM lays emphasis on the separation of the two classes as conventional SVM. Fig. 2 shows how the resulted hypersphere changes when different γ are used.

By introducing multipliers α_i for the inequality constraints, we obtain the duel problem of Eq. (3) as

$$\begin{aligned} \max_{\alpha_i} \quad &: \quad -\frac{1}{\gamma}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N}\alpha_i y_i x_i^T x_i \\ \text{subject to} \quad &: \quad \sum_{i=1}^{N}\alpha_i y_i = \gamma, \quad \sum_{i=1}^{N}\alpha_i = 1, \\ & \quad\ \alpha_i \geq 0, \ \text{ for } i = 1, \dots, N. \end{aligned}$$

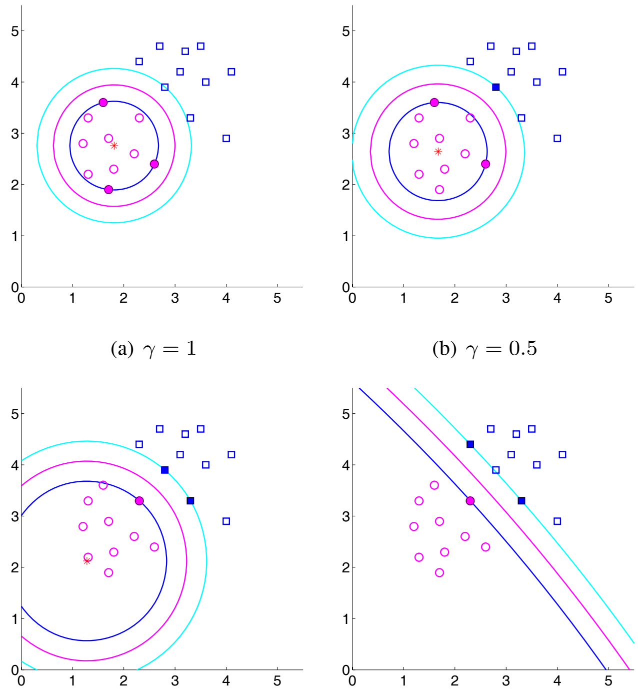
It can be shown that the center a of the hypersphere is a linear combination of the training samples as $a = \sum_{i=1}^{N}\alpha_i y_i x_i/\gamma$. Similar to SVM, the optimization of (4) is a quadratic programming problem, and as a result many α_i are zeros. In other words, the center are determined by a few training samples and we also call them support vectors. It is easy to check that the parameter γ should satisfy $0 \leq \gamma \leq 1$ for (4) to have feasible solutions.

When the samples cannot be separated by a sphere, we have to allow some negative samples inside the enclosing hypersphere and/or some positive samples outside the excluding hypersphere. In analogy to what SVM does in the situation, we introduce slack variables $\xi_i \geq 0$ and Eq. (3) becomes

$$\begin{aligned} \min_{a, R^2, \Delta R^2, \xi_i} \quad &: \quad \gamma R^2 - \Delta R^2 + C\sum_{i=1}^{N}\xi_i \\ \text{subject to} \quad &: \quad ||a - x_i||^2 \leq R^2 - \Delta R^2 + \xi_i \ \text{ for } y_i = 1, \\ & \quad\ ||a - x_i||^2 \geq R^2 + \Delta R^2 - \xi_i \ \text{ for } y_i = -1, \\ & \quad\ \Delta R^2 \geq 0. \end{aligned} \tag{4}$$

Note one additional constraint $\Delta R^2 \geq 0$ is added in Eq. (4) to force the enclosing ball to be inside of excluding ball, which is not assured in the non-separable case.

The dual problem of (4) is very similar to that of (4). The differences lie in the the upper bound of α_i and the



(a) $\gamma = 1$ (b) $\gamma = 0.5$

(c) $\gamma = 0.2$ (d) $\gamma = 0.01$

Fig. 2. The effect of parameter γ on the resulted sphere of MEMEM. The center is depicted as a read star. The circles in blue, magenta and cyan represent the enclosing ball $B(a, R_1)$, the decision boundary $B(a, R)$ and the excludding ball $B(a, R_2)$ respectively.

additional multiplier β associated with the constraint $\Delta R^2 \geq 0$.

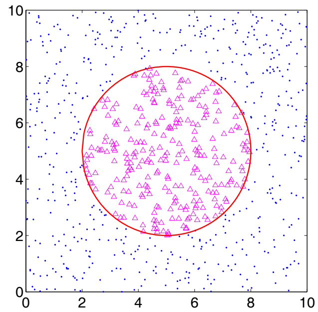
$$
\begin{aligned}
\max_{\alpha_i, \beta} \quad &: \quad -\frac{1}{\gamma}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^{N}\alpha_i y_i x_i^T x_i \\
\text{subject to} \quad &: \quad \sum_{i=1}^{N}\alpha_i y_i = \gamma, \quad \sum_{i=1}^{N}\alpha_i - \beta = 1, \\
&\quad\;\; \beta \geq 0, \quad C \geq \alpha_i \geq 0, \ \text{for } i = 1, \ldots, N.
\end{aligned}
$$

It can be shown that by choosing $\gamma = 1$, Eq. (4) becomes the soft-margin SVDD, and therefore with or without the negative samples, MEMEM has SVDD as one special case when MEMEM only deals with the task of description.

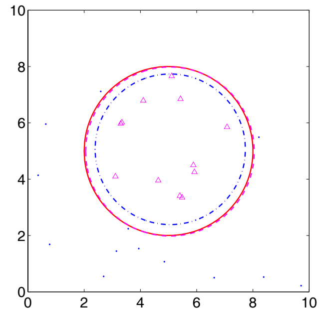
III. Experimental Results

A. Synthetic Data

Before conducting experiments on real-life data, we try MEMEM on a 2-D synthetic data set to illustrate the performance. The data are uniformly distributed in a square as shown in Fig. 3(a). The triangles and the dots represent the target class and the non-target class, which are the samples inside and outside the ball with radius 3 centered at (5, 5), respectively. First, we randomly generate 26 training samples with 13 from each class. The samples generated in one simulation are depicted in Fig. 3(b), where one can see that the training samples are



(a)



(b)

Fig. 3. (a) The class distribution of the synthetic data. The true boundary is the circle with radius 3 centered at (5, 5). (b) The true boundary (solid line in red), the boundary produced by SVDD (the dot-dashed line in blue), and the boundary produced by MEMEM (the dashed line in magenta).

quite sparse. The decision boundaries produced by SVDD and MEMEM ($\gamma = 0.5$), which are averaged over 100 simulations, are drawn along with the true boundary in Fig. 3(b). Evidentally, MEMEM outperforms SVDD.

Unbalanced training data are also generated. We test two cases: (1) $\eta = 5/2$; and (2) $\eta = 2/5$, where η denotes the ratio of the number of training samples in the target class and the non-target class. The average center and radius of the ball produced by SVDD and MEMEM are listed in Table I. Both SVDD and MEMEM find the center very well, and we believe it is because the shape of the true boundary, which is a circle, fits the model chosen by both method. The major difference, however, is the radius. Due to its descriptive nature, the emphasis of SVDD is to make the ball as small as possible and therefore it produces a much smaller ball than the true one. MEMEM, on the other hand, also considers the separation between the classes. As a result, the resulted boundary is very close to the true boundary even when the samples are scarce.

To test how MEMEM performs when the training set is relatively abundant, we conduct the second series of simulations by utilizing 100 training samples. Again, we run both balanced and unbalanced data, and list the results in Table II. As one can see, with more training data, the ball found by SVDD is expanded and more closer to the true boundary. However, it is still outperformed by MEMEM, which almost produced the perfect decision boundary.

B. Medical Diagnosis Data

For the real-life data, we investigate the performance of MEMEM on the Biomed data set [9] from the Statlib data archive [10], which was collected in a study that aims at developing screening methods to identify carriers of a rare genetic disorder. Excluding 15 observations that are not usable due to missing values, this biomedical data set contains 194 observations with 127 normal samples from healthy patients and 67 abnormal samples from disease carriers. Each observation has 4 attributes corresponding to measurements taken on blood samples.

First, we train MEMEM with balanced data ($\eta = 1$). In total, 100 samples are randomly chosen with 50 each from normal and abnormal classes. We are left with 77 normal observations and 17 abnormal observations, which

TABLE I

THE COMPARISON OF THE CENTER a AND RADIUS R OBTAINED BY SVDD AND MEMEM WHEN 25 TRAINING SAMPLES ARE USED.

	a		R	
	SVDD	MEMEM	SVD	MEMEM
$\eta = 5:2$	(4.99, 4.92)	(4.99, 4.92)	2.74	3.04
$\eta = 1:1$	(5.06, 5.01)	(5.05, 4.98)	2.67	3.00
$\eta = 2:5$	(5.04, 4.99)	(5.05, 4.97)	2.50	2.98

TABLE II

THE COMPARISON OF THE CENTER a AND RADIUS R OBTAINED BY SVDD AND MEMEM WHEN 100 TRAINING SAMPLES ARE USED.

	a		R	
	SVDD	MEMEM	SVD	MEMEM
$\eta = 5:2$	(4.99, 5.01)	(4.99, 5.01)	2.94	3.02
$\eta = 1:1$	(4.99, 4.99)	(4.99, 5.00)	2.93	3.01
$\eta = 2:5$	(4.99, 5.01)	(4.99, 5.01)	2.95	3.01

are used as the test samples. RBF, which is reported as a good choice for this data set [6], is adopted as the kernel function. The generalization accuracy is estimated by using different parameters in Eq. (4). More specifically, we try $C = [0.1, 0.2, \ldots, 1]$ and $\gamma = [0.05, 0.1, 0.2, \ldots, 1]$, which provides $10 \times 11 = 110$ combinations, to find the best pair (C, γ) that yields the highest classification accuracy on the test samples.

SVDD and SVM are also employed to make the performance comparison. Different from MEMEM, SVM and SVDD have only one parameter to tune, which shares the same notation as C in the formulation. For SVM, we test 13 choices as $C = 10^{-1} \cdot [2^0, 2^1, \ldots, 2^{12}]$. SVDD, however, requires the parameter to be $0 \leq C \leq 1$ and thus we vary C from 0 to 1 with a step size of 0.1. Again, each machine is trained for every choice of C and the highest classification accuracy is selected for the performance comparison.

Other scenarios that test how MEMEM performs on unbalanced data set are also simulated. The description of the data shows that because the disease is rare, there are only a few carriers of the disease from whom data are available. For this reason, we always have the training set contain more normal observations than abnormal ones to make the simulations practical. Three cases are experimented and the results are listed in Table III. As one can see, for all the scenarios simulated, MEMEM consistently performs the best or the comparably best among the three learning machines. When $\eta = 1$, SVM and MEMEM yield comparable performance. This is because with ample normal and abnormal samples, this problem can be approached by a traditional binary classifier. However, when $\eta = 8:1$ for which the number of abnormal samples are relatively few, SVM, which focuses on discrimination between classes, is outperformed by both SVDD and MEMEM. Actually, the accuracy yielded by MEMEM, which is 81.4%, is significantly higher than that of the other two machines, which are 76.3% and 72.3% respectively.

TABLE III

THE COMPARISON OF SVDD, MEMEM AND SVM ON THE BIOMED DATA SET TRAINED BY 100 TRAINING SAMPLES. THE OPTIMAL PARAMETERS ARE LISTED IN THE PARENTHESIS.

η	# of training samples		SVDD	MEMEM	SVM
	normal	abnormal	(C)	(C, γ)	(C)
$8:1$	89	11	76.3% (0.6)	81.4% $(0.1, 0.7)$	72.3% (12.8)
$4:1$	80	20	81.4% (0.8)	88.7% $(0.8, 0.3)$	82.5% (102.4)
$2:1$	67	33	79.4% (0.1)	89.7% $(0.9, 0.3)$	90.1% (6.4)
$1:1$	50	50	82.5% (0.6)	92.8% $(0.05, 0.1)$	91.8% (204.8)

IV. CONCLUSIONS

MEMEM, a new algorithm for robust classification is proposed. Rooted in one-classification, MEMEM retains the descriptive capability of the positive data by enclosing them with a small hypersphere and is able to reject the uncharacteristic patterns. In the meantime, MEMEM exploits the discriminating information provided by the negative samples and therefore is more robust for generalization.

One of the key parameters of MEMEM is $\gamma \in [0\ 1]$, which controls the relative weight of the description and discrimination factors considered in the formulation. During the experiments, the best γ is chosen by trying different values. For the synthetic data set, the results are similar for a large range of γ while for the biomed data set the choice of γ has much more influence on the performance of MEMEM. One future research direction is to study in deep the relationship between γ, the nature of the data, and the resulted boundary such that the optimal one can be found more efficiently.

REFERENCES

[1] Y.Q. Chen, X.S. Zhou, and T.S. Huang, One-Class SVM for Learning in Image Retrieval, *Proceedings of IEEE International Conference on Image Processing*, vol. 1, pp. 34-37, 2001.

[2] M. Gori and F. Scarselli, Are Multilayer Perceptrons Adequate for Pattern Recognition and Verification? *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, 1998.

[3] M. Moya, M. Koch, and L. Hostetler, One-Class Classifier Networks for Target Recognition Applications, *Proceedings of World Congress on Neural Networks*, pp. 797-801, 1993.

[4] M.J. Tax and R.P.W. Duin, Support Vector Domain Description, *Pattern Recognition Letters*, vol. 20, no. 11-13, pp. 1191-1199, 1999.

[5] B. Schölkopf, J.C. Platt, J. Shawe-Taylor, A.J. Smola, and R.C. Williamson, Estimating the Support of a High Dimensional Distribution, *Neural Computation*, vol. 13, no. 7, pp. 1443-1471,1999.

[6] C. Campbell and K.P. Bennett, *A Linear Programming Approach to Novelty Detection*, *Advances in Neural Information Processing Systems*, vol. 13, MIT press, Cambridege, MA, 2001, pp. 395-401.

[7] K. Goh, E.Y. Chang, and B. Li, Using One-class and Two-Class SVMs for Multiclass Image Annotation, *IEEE Tran. on Knowledge and Data Engineering*, vol. 17, no. 10, pp. 1333-1346, 2005.

[8] C. Yuan and D. Casasent, A Novel Support Vector Classifier with Better Rejection Performance, *Proceedings of CVPR*, vol. 1, pp. 419-424, 2003.

[9] L.H. Cox, M.Johnson, and K.Kafadar, Exposition of Statistical Graphics Technology, *ASA Proceedings of the Statistical Computation Section*, pp. 55-56, 1982.

[10] http://lib.stat.cmu.edu/datasets.