Learning the Boundary of One-Class-Classifier Globally and Locally

Aimin Feng, Bin Chen, XuejunLiu
Computer Science & Technology College
Nanjing University of Aeronautics & Astronautics
210016, Nanjing, P.R. China
{amfeng, B.Chen, xuejun.liu}@nuaa.edu.cn

Bin Chen
Dept. of Computer Science
Yangzhou University
225009, Yangzhou, P.R. China

***Abstract*—The One class classification problem aims to distinguish a target class from outliers. Two popular algorithms, One-Class SVM (OCSVM) and Single-Class MPM (SCMPM), solve this problem by finding a hyperplane with the maximum distance to the origin. Their essential difference is that OCSVM focuses on the Support Vectors (SV) in a local manner while SCMPM emphasizes the whole data's distribution using global information. In fact, these two seemingly different yet complementary characteristics are all important prior knowledge for the One-Class-Classifier (OCC) design. In this paper, we propose a novel OCC called Global & Local (GLocal) OCC, which incorporates the global and local information in a unified framework. Through embedding the samples' distribution information into the original OCSVM, the GLocal OCC provides a general way to extend the present SVM algorithm to consider global information. Moreover, the optimization problem of the GLocal OCC can be solved using the standard SVM approach similar to OCSVM, and preserves all the advantages of SVM. Experiment results on benchmark data sets show that the GLocal OCC really has better generalization compared with OCSVM and SCMPM.**

***Keywords*—One-Class-Classifier, Globality, Locality, Support Vector, Sparsity, Quadratic Programming**

I. INTRODUCTION *(HEADING 1)*

In One-Class classification problems, usually, only one class of data is available, but others are too expensive to acquire or too difficult to characterize. The available class is called the target class or normal patterns, while all others not in this class are defined as outliers or abnormal patterns. These classification tasks can be found in many real-world scenarios like machine faulty diagnosis, network intrusion detection and document classification etc. Originating from various applications, one-class classification is also referred to as domain description, novelty detection, or concept learning [1].

To solve the one-class classification problem, an extreme approach is to estimate the probability density function (pdf) of the data in the target class [2]. But since Vapnik et al. proposed a principle [3] that "*never to solve a problem that is more general than the one we actually need to solve*",an alternative solution, the domain-based approach [1,4], has become the main method to solve one-class classification problem. Through finding a boundary to enclose the target class appropriately, the domain based method minimizes the volume of the target class domain by geometric shapes such as hyperplane or hypersphere.

As the state-of- the-art SVM introduced in one class, One-Class SVM (OCSVM) [5] uses a hyperplane to separate the target samples from the origin with maximal margin. In nonlinear case, the kernel trick can be used by first implicitly mapping the training data into a higher dimensional feature space and then using the linear method. Moreover, when Gaussian kernels are used, the OCSVM is equivalent to that of Support Vector Data Description (SVDD) [6], which describes the target data domain by finding the minimum hyperspere. Computationally, the above two methods both lead to quadratic programming (QP). For further reducing the computational cost, a linear programming algorithm is proposed on the basis of OCSVM [7].

Using the dual formulation to solve the convex optimization as SVM usually does, OCSVM has the advantages including the sparse solution, the global optimum and the large margin [8]. However, it only employs the Support Vectors（SV） but neglects the contribution of the whole samples' distribution to the margin. Such a local learning characteristic is more likely to lose the global information of the whole data so that it has the dangerous to make its boundary sub-optimal. More recently, some algorithms such as Single-Class Mini-Max Probability Machine (SCMPM) [9], Mahalanobis One-Class SVM (MOCSVM) [10], Minimum Volume Enclosing Ellipsoidal (MVEE) [4] and Minimum Volume Covering Ellipsoid (MVCE) [11], have, in fact, given justice that the utility of the global information in data is vital for designing a classifier since it can lead to more powerful generalization.

For the sake of simplicity, here we only focus on the above two algorithms which use the hyperplanes model, i.e, SCMPM and MOCSVM. SCMPM maximizes the Mahalanobis Distance (MD) of the hyperplane to the origin instead of the Euclidean Distance (ED) in OCSVM. Given only the mean and covariance matrix as the global information of the data distribution, this model exploits the worst-case probability of the target data falling inside the positive half space. Inspired by SCMPM, Tsang et al. proposed MOCSVM which also utilizes MD to improve the OCSVM's performance. To alleviate the

This work is supported by the National Nature Science foundation of China, No.60603029 and No.60703016.

estimation error for the global issues of the first and second order moments, the above models both adopted a uncertainty model named robust estimation. Unfortunately, this robust estimation may be inaccurate but seems to need the local characteristic of the specific data points. Furthermore, without using the dual theory but solving the primal problem directly, SCMPM loses the sparsity derived from the KKT conditions. In addition, the Second Order Cone Programming (SOCP) optimization is computationally time-consuming compared to the QP optimization used in OCSVM.

Inspired by the above analyses on the existing One-Class-Classifier (OCC) algorithms, we propose a Global & Local (GLocal) OCC by incorporating the global and local information in an integrative framework. Through embedding the samples' distribution information into the original OCSVM, the GLocal OCC provides a general way to extend the classical SVM algorithm. In such a way, GLocal OCC is able to preserve all the advantages of SVM including the global optimality, the sparse solution and the large margin. Moreover, the optimization of the GLocal OCC can be solved using the standard SVM approach similar to OCSVM rather than SOCP as in SCMPM.

The rest of this paper is organized as follows: Section 2 briefly introduces the OCSVM and SCMPM as the related works. Section 3 presents our GLocal OCC and its kernelized version. Section 4 gives the experimental results on benchmark datasets. Finally, some conclusions are drawn in Section 5.

II. RELATED WORKS

A. One-Class SVM(OCSVM)

Given a set of patterns $\{x_1, x_2, \cdots, x_n\}$, OCSVM [5] tries to find the hyperplane which separates most samples with the maximal margin from the origin by solving

$$\begin{aligned} &\min_{\mathbf{w},\zeta,\rho} \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} - \rho + \frac{1}{\nu n}\sum_{i=1}^{n}\zeta_i \\ &s.t.\ \mathbf{w}^{\mathrm{T}}x_i \geq \rho - \zeta_i, \zeta_i \geq 0, i = 1,...,n \end{aligned} \tag{1}$$

$\mathbf{w}^{\mathrm{T}}x_i = \rho$ is the desired hyperplane and the margin from the origin to the hyperplane is $\rho / \sqrt{\mathbf{w}^{\mathrm{T}}\mathbf{w}}$. $\nu \in (0,1)$ is the parameter which characterizes the fraction of support vectors and outliers. ζ_i is the slack variables used to penalize the samples lying on the negative half space.

The corresponding *Wolfe dual form* of (1) is

$$\begin{aligned} &\min_{\alpha} = \frac{1}{2}\alpha^{T}\mathbf{X}^{\mathrm{T}}\mathbf{X}\alpha \\ &s.t.\ \ \alpha^{T}\mathbf{1} = 1,\ 0 \leq \alpha \leq \frac{1}{vl}\mathbf{1} \end{aligned} \tag{2}$$

where $\alpha = [\alpha_1, \cdots, \alpha_n]^{\mathrm{T}}$, $\mathbf{1} = [1, \cdots, 1]^{\mathrm{T}}$.

The dual formulation has two outstanding properties: first, the solution is sparse, which means that only the training samples that lie on the surface of the optimal hyperplane have their corresponding nonzero α_i and are called Support Vectors (SV). Second, it is possible to use the mercer kernel $k(\cdot,\cdot)$ to replace the dot product $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ such that it implicitly maps the input samples into the higher-dimensional feature space. According to Cover's theory [2], the nonlinear problem in the data space can be more likely linearly solved in the mapped space.

For a given new test point z, the following decision function determines whether the point belongs to the target class or outliers:

$$f(x) = \mathrm{sgn}\left[\alpha^{T}\mathbf{X}^{\mathrm{T}}z - \rho\right] = \begin{cases} 1 & \text{target class} \\ -1 & \text{outlier} \end{cases} \tag{3}$$

B. Single-Class MPM (SCMPM)

Similar to the OCSVM, SCMPM finds the smallest half-space $Q(\mathbf{a},b) = \{x | \mathbf{a}'x \geq b\}$ for the normal patterns by maximizing the Mahalanobis distance from hyperplane to the origin. Given only the mean $\bar{\mathbf{x}}$ and the covariance matrix $\mathbf{\Sigma}$ of a distribution and without further assumptions on the data, the SCMPM minimizes the worst-case probability of a data pattern falling inside Q. Hence, for a given $\alpha \in (0,1)$, this leads to the following constrained optimization problem:

$$\max \frac{b}{\sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{a}}} \quad s.t. \quad \inf\ \mathrm{P}(\mathbf{a}^{\mathrm{T}}x \geq \mathrm{b}) \geq \alpha \tag{4}$$

Without loss of generality, we set $b = 1$ in (4). In addition, by using the generalized Chebychev inequality, the constraint in (4) is the same as the ones in the following optimization problem:

$$\begin{aligned} &\min \quad \sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{a}} \\ &s.t. \quad (\mathbf{a}^{\mathrm{T}}\bar{\mathbf{x}} - 1) \geq \kappa(\alpha)\sqrt{\mathbf{a}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{a}} \end{aligned} \tag{5}$$

where $\kappa(\alpha) = \sqrt{\alpha/1-\alpha}$.

The above optimization problem needs to be solved by SOCP, which is computationally time-consuming compared to the QP solver. Moreover, the solution loses the sparseness property. In addition, the SCMPM heavily depends on the mean and the covariance which are estimated from the data. To improve the robustness, a uncertainty model is adopted for the estimation as the following:

$$\{(\bar{\mathbf{x}}, \mathbf{\Sigma}) : (\bar{\mathbf{x}} - \bar{\mathbf{x}}^{\mathbf{0}})^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\bar{\mathbf{x}} - \bar{\mathbf{x}}^{\mathbf{0}})\} \leq \delta^2, \left\|\mathbf{\Sigma} - \mathbf{\Sigma}^{\mathbf{0}}\right\|_F \leq \gamma \tag{6}$$

Where $\bar{\mathbf{x}}^{\mathbf{0}}$ and $\mathbf{\Sigma}^{\mathbf{0}}$ are the nominal estimates of $\bar{\mathbf{x}}$ and $\mathbf{\Sigma}$, $\|\cdot\|_F$ denotes the Frobenious norm and $\delta, \gamma \geq 0$ are the corresponding uncertainties.

It is worth noting that SCMPM can also obtain the nonlinear region by using the kernel trick even without exploiting the dual theory. For the reason that the mean and the covariance matrix can be denoted by the samples in first and second order moments, the optimum will lie in the span of the sample data so that the objective function and the constraints of

(5) can be expressed in terms of inner products. For more the detailed derivation, please refer to [12].

III. GLOCAL ONE-CLASS-CLASSIFIER

From the formulation of the related work, we draw the idea that OCSVM stresses much more on the local information of the support vectors but pays less attention to the whole data's structure. Conversely, SCMPM emphasizes more on the global information by using sample's mean and covariance while neglecting individual data's effect on the boundary. Although these two algorithms compete for each other in solving one-class classification, they both indeed lose some useful information during constructing classifier. Motivated by unifying the global and the local issues into an integrated framework, we proposed a novel OCC called Global&Local One-Class-Classifier (GLocal OCC). Through incorporating the covariance matrix into the original OCSVM, the GLocal OCC takes into account the sample structure at the same time maximizing the margin to the hyperplane . In the following sections, we will discuss the linear and kernel versions of GLocal OCC respectively.

A. Linear GLocal OCC

Since the covariance matrix usually expresses global information and the SVM framework is relatively convenient to solve with QP, we arbitrarily embed the covariance matrix $\mathbf{\Sigma}$ of the whole data into the original OCSVM formulation. Coinciding with (1) of OCSVM, here we describe the soft margin objective function as the following:

$$\begin{aligned} &\min_{\mathbf{w},\zeta,\rho} \ \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}+\frac{1}{2}\lambda\mathbf{w}^{\mathrm{T}}\mathbf{\Sigma}\mathbf{w}-\rho+\frac{1}{vn}\sum_{i=1}^{n}\zeta_i \\ &s.t. \ \mathbf{w}^{\mathrm{T}}x_i \geq \rho-\zeta_i \quad \zeta_i \geq 0 \end{aligned} \tag{7}$$

where $\lambda \geq 0$ is the *trade-off parameter* that regulates the balance between the new term and the original formulation. Here the $\mathbf{\Sigma}$ denotes the covariance matrix of the samples in order to represent the ***global issue*** of the input data, while the other items are the same as OCSVM which try to find support vectors referred to as the local characteristic of the training data.

Transforming the primal problem into its corresponding dual one, we have:

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\alpha^{T}\mathbf{X}^{\mathrm{T}}\left(\mathbf{I}+\lambda\mathbf{\Sigma}\right)^{-1}\mathbf{X}\alpha \\ &s.t. \ \alpha^{T}\mathbf{1}=1, \quad 0\leq\alpha\leq\frac{1}{vn}\mathbf{1} \end{aligned} \tag{8}$$

Compared to the dual (2) of OCSVM, the above dual (8) is not in the input space defined by the inner product (x_i, x_j), but replaced by $\mathbf{X}^{\mathrm{T}}\left(\mathbf{I}+\lambda\mathbf{\Sigma}\right)^{-1}\mathbf{X}$, which is equivalent to mapping the samples into a new feature space by $(\mathbf{I}+\lambda\mathbf{\Sigma})^{-\frac{1}{2}}$. That is, GLocal OCC finds a hyperplane that is in a different space from the one in the OCSVM. When this hyperplane is mapped back to the input space, it becomes a nonlinear boundary which is undoubtedly with more separable ability than the linear hyperplane.

It is worth noting that solving process for GLocal is almost the same as the OCSVM, thus off-the-shelf QP solver or some decomposition methods such as SMO[13] can be exploited even without any modification.

The decision function is described as:

$$f(\mathbf{x}) = \mathrm{sgn}\left[\alpha^{T}\mathbf{X}^{\mathrm{T}}(\mathbf{I}+\lambda\mathbf{\Sigma})^{-1}z-\rho\right] \tag{9}$$

From the above process, we can see that the GLocal OCC is similar to the OCSVM no matter in primal, in dual and in decision function. If without considering the covariance information, (7), (8) and (9) would be reduced to (1), (2) and (3) respectively. In other words, the GLocal OCC is the generalization of the OCSVM except that it takes into account the global information of the data's distribution.

For further explanation, Fig. 1 displays the geometric interpretation of GLocal OCC. In this figure, the target samples distribute in an ellipsoid. Each point therefore has additional global information of the whole data. For simplicity, Fig. 1 only shows some points illustrated by the dash-dotted ellipsoids with the same shape. However, OCSVM doesn't consider this global issue: Its decision hyperplane H1 locates on the points nearest to the origin which is indicated by the SVs in the Fig 1. In comparison, GLocal constructs its decision plane H2 according to both the global and local information: The GLocal hyperplane is adjusted by the tangent hyperplane of the dash-dotted ellipsoids centered at the support vectors (the local information) and the covariance of the members in the target class (the global information).

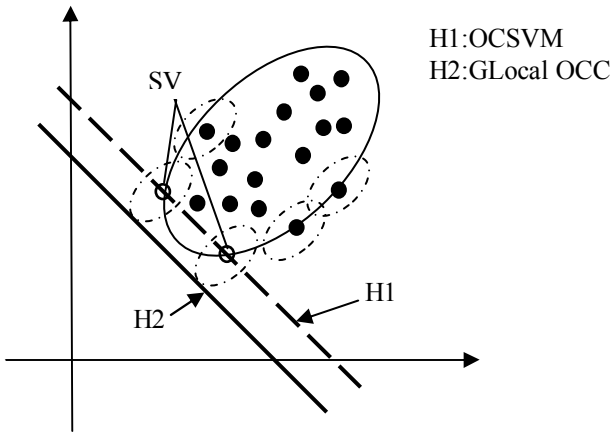

Figure 1. A geometric interpretation of GLocal OCC.

B. Kernel GLocal OCC

Here we omit the formulas of kernel GLocal OCC but still utilize the linear version (7) (8) and (9). In order to utilize the kernel trick in the dual form, all the terms of $\mathbf{X}^{\mathrm{T}}\left(\mathbf{I}+\lambda\mathbf{\Sigma}\right)^{-1}\mathbf{X}$ need to be denoted by the inner product. For this reason, the $\mathbf{\Sigma}$ is described as

$$\mathbf{\Sigma}=\frac{1}{\mathbf{n}}\mathbf{X}\mathbf{X}^{\mathrm{T}}-\frac{1}{\mathbf{n}^{2}}\mathbf{X}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}=\frac{1}{\mathbf{n}}\mathbf{X}\mathbf{H}\mathbf{X}^{\mathrm{T}} \tag{10}$$

Here $\mathbf{X}$ is the vector of the input patterns. Denote $\mathbf{H} = (\mathbf{I} - \frac{1}{\mathbf{n}}\mathbf{1}\mathbf{1}^{\mathbf{T}})$, where $\mathbf{I}$ is the identity matrix. So by using the following Woodbury formula [14]

$$(\mathbf{A}+\mathbf{BC})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I}+\mathbf{C}\mathbf{A}^{-1}\mathbf{B})^{-1}\mathbf{C}\mathbf{A}^{-1}$$

And using the properties of $\mathbf{HH} = \mathbf{H}$ and $\mathbf{H} = \mathbf{H}^{\mathbf{T}}$, we obtain

$$(\mathbf{I}+\lambda\boldsymbol{\Sigma})^{-1} = \mathbf{I} - \frac{\lambda}{n}\mathbf{X}\mathbf{H}(\mathbf{I}+\frac{\lambda}{n}\mathbf{H}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{H})^{-1}\mathbf{H}\mathbf{X}^{\mathbf{T}} \qquad (11)$$

By adopting the kernel trick, (8) then becomes:

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\alpha^{T}(\mathbf{K} - \frac{\lambda}{n}\mathbf{K}\mathbf{H}(\mathbf{I}+\frac{\lambda}{n}\mathbf{H}\mathbf{K}\mathbf{H})^{-1}\mathbf{H}\mathbf{K})\alpha \\ &s.t. \ \ \alpha^{T}\mathbf{1} = 1, \quad 0 \le \alpha \le \frac{1}{vn}\mathbf{1} \end{aligned} \qquad (12)$$

where $\mathbf{K} = \mathbf{X}^{\mathbf{T}}\mathbf{X}$ is the kernel matrix. This is again a standard QP. Moreover, when $\mathbf{K}$ is invertible, by using the Woodbury formula, (12) can be further simplified (the detailed derivation is in the Appendix):

$$\begin{aligned} &\min_{\alpha} \ \frac{1}{2}\alpha^{T}(\mathbf{K}^{-1} + \frac{\lambda}{n}\mathbf{H})^{-1}\alpha \\ &s.t. \ \ \alpha^{T}\mathbf{1} = 1, \quad 0 \le \alpha \le \frac{1}{vn}\mathbf{1} \end{aligned} \qquad (13)$$

The decision function (9) is also changed to:

$$f(\mathbf{x}) = \mathbf{sgn}\left[\alpha^{\mathbf{T}}(\tilde{\mathbf{K}} - \frac{\lambda}{n}\mathbf{K}\mathbf{H}(\mathbf{I}+\frac{\lambda}{n}\mathbf{H}\mathbf{K}\mathbf{H})^{-1}\mathbf{H}\tilde{\mathbf{K}} - \rho\right] \qquad (14)$$

where $\tilde{\mathrm{K}}$ represents the kernel matrix between training and testing samples. Since it is not a square matrix, the decision function can not be further simplified as (13).

IV. EXPERIMENT

Since there are hardly one-class benchmark data sets, we compared the performances of GLocal OCC with OCSVM and SCMPM on seven binary class from the UCI machine learning repository as the OCC usually does [12][15]. Table I lists all these data sets which are divided into three groups by dimension from low to high. For each data set, we follow the step in [15] to take the larger class as normal data and the other as outliers. We randomly sample 80% of the normal patterns for training, and the remaining 20% of the normal patterns and all the outliers are used for testing.

TABLE I. DATA SETS USED IN EXPERIMENT

Group	dataset	dimension	total	training	testing	
				normal	normal	outlier
Low	Biomed	5	194	102	25	67
	Breast cancer	9	699	367	91	241
	Heart	13	303	123	41	139
Medium	Import	25	159	71	17	71
	Ionosphere	34	351	180	45	126
High	Sonar	60	208	89	22	97
	Arrhythmia	278	420	190	47	183

In the experiment, we use a Gaussian kernel

$$K(x,y) = e^{-\|x-y\|^2/\sigma^2}$$

where the kernel parameter σ is tuned by the grid search, so is the hyperparamete λ in GLocal OCC. Set $\nu = 0.1$ both in OCSVM and GLocal OCC. For the α in SCMPM, although in theory it should be set near to $1-\nu = 0.9$, in our experiment the results can not compete with OCSVM until it decreases to 0.6. This value is coincident with the experiment result reported in [16], which concludes SCMPM has comparative performance when $\alpha \in [0.6, 0.8]$. In addition, we set $\rho = 0.01$ to the plug-in estimate of the covariance matrix as in [12].

In one-class classification, according to its true labels and classified results, there are four possible cases listed in Table II [1] called confusion matrix.

There are two kinds of errors denoted by italic in Table II called False Positive (FP) and False Negative (FN). In experiment, we adopt FP/FN together with Balanced Loss (BL) [15] to evaluate the results of the algorithms, here BL=$(\mathrm{FP}+\mathrm{FN})/2$. Obviously, the lower of the above criteria, the better performance of the algorithms. Experiment results are shown in Table III. To reduce statistical variability, average results of 10 repetitions are reported. The italic and bold font denotes the best result of each data set according to the Balance Loss.

TABLE II. ALL CASES FOR OCC

Classified label	True label	
	Target	outlier
target	True Positive(TP)	*False Negative(FN)*
outlier	*False Positive(FP)*	True Negative(TN)

TABLE III. TEST RESULTS ON THE UCI DATA

Data Sets	OCSVM			SCMPM			GLocal OCC		
	FP	FN	BL	FP	FN	BL	FP	FN	BL
Biomed	0.1418	0.1520	0.1469	0.2119	0.0880	0.1500	***0.1313***	***0.1120***	***0.1217***
Breast cancer	0.0228	0.0604	0.0416	0.0622	0.0319	0.0471	***0.0274***	***0.0473***	***0.0373***
Heart	0.5424	0.2781	0.4103	0.7691	0.1531	0.4611	***0.5165***	***0.2531***	***0.3848***
Import	0.1775	0.3353	0.2564	0.2338	0.2294	0.2316	***0.1817***	***0.2294***	***0.2056***
Ionosphere	0.0317	0.1556	0.0937	0.1183	0.0689	0.0936	***0.0325***	***0.1156***	***0.0740***
Sonar	0.1165	0.6591	0.3878	0.6567	0.1000	0.3784	***0.1351***	***0.5409***	***0.3380***
Arrhythmia	0.4612	0.1191	0.2902	0.4732	0.0064	0.2941	***0.3798***	***0.1596***	***0.2697***

In Table III, for the comparison of BL, we have noticed that GLocal OCC is better than the other two algorithms in all of the seven datasets consistently, no matter in low, medium and high dimensions. Particularly in Sonar, the BL decreases nearly 5~6% compared to OCSVM and SCMPM. For other data sets, GLocal OCC obtains at least 2~3% better than the others. These results sufficiently prove that it is more reasonable of considering both the global and local information than only taking one into account.

According to further analysis, we found the small values of BL obtained by GLocal OCC are mainly caused by the lower values of FN compared with OCSVM (except Arrhythmia). It is reasonable since GLocal OCC considers the target data's distribution in finding its decision boundary just as shown in Fig. 1, this hyperplane can undoubtedly cover more target data. For the exceptional case of Arrhythmia, it may be due to that its training sample is too small (190) compared with its high dimension (278). We also notice that this hyperplane possibly leads to large FP since its enlarged boundary has the risk to include the space of the outliers. This dilemma can be reduced if the negative samples are supplied in training. In one-class classification, we have to burden this risk. However, the increase of FP is usually slower than the decrease of FN, so we can get these satisfactory results. This further proves that it is reasonable to take into account the data's distribution in OCSVM.

For the SCMPM, since the loose possibility α is set for the sake of comparative BL, FP/FN shows great unbalance, FP is almost much bigger than FN except Import. On the one hand, it shows that SCMPM is really too cautious to label samples as outlying and therefore has a high FP rate[15]; On the other hand, this deeply proves it is not a valid way for SCMPM to only focus on global information without caring about local characteristic of the individual data.

V. CONCLUSION

In this paper, we proposed a novel classifier called GLocal OCC. Through embedding the covariance matrix into the original OCSVM, GLocal OCC shows better generalization proved by the experiment results on benchmark datasets. At the same time, this new model provides a general method to incorporate the global information into the SVM framework with only local characteristic. Since it is also a QP problem, the standard SVM approach can be employed to solve the optimization similar to OCSVM. In addition, GLocal OCC still keep all the advantages of OCSVM such as the global optimality, the sparse solution and the large margin. In future work, inspired by sTructure OCC (TOCC) [16] which further considers the data distribution in delicate granularity, we will extend our work under present framework in finer clusters within the target class.

ACKNOWLEDGMENT

The authors thank the constructive discussion with Professor S. Chen and Mrs. Hui Xue of the Parnec group.

REFERENCES

[1] D.M.J. Tax, "One-class classification: Concept-learning in the absence of counter-examples," Delft University of Technology. 2001

[2] R.O. Duda, P.E. Hart and D.G. Stork, "Pattern Classification," 2nd Edition, New York: John Wiley & Sons, 2001.

[3] V. Vapnik, "Statistical Learning Theory," Addison-Wiley, 1998.

[4] P. Juszczak, "Learning to recognise: A study on one-class classiifcation and active learning," Delft University of Technology, 2006.

[5] B. Schölkopf, J.C. Platt, and J. Shawe-Taylor, "Estimating the support of a high-dimensional distribution," Neural Computation, 2001, vol. 13, No.7, pp. 1443-1471.

[6] D.M.J. Tax and R.P.W. Duin, "Support Vector Data Description," Machine Learning, 2004. vol. 54, No. 1, pp. 45-66.

[7] C. Campbell and K. P. Bennett, "A Linear Programming Approach to Novelty Detection," Advances in Neural Information Processing Systems, Cambridge, MIT Press, 2001.

[8] J. Shawe-Taylor and N. Cristianini, "Kernel Methods for Pattern Analysis," Cambridge University Press, 2004.

[9] G. Lanckriet, L. E. Ghaoui and M. Jordan, "Robust novelty detection with single-class MPM," Advances in Neural Information Processing Systems, 2002.

[10] I, W. Tsang, J. T. Kwok and S. Li, "Learning the kernel in Mahalanobis one-class support vector machines," Proceedings of the International Joint Conference on Neural Networks, pp.1169-1175, Vancouver, Canada, July 2006.

[11] D. Alexander, D.T. Bie, C. Harris and Shawe-Taylor, "The Minimum volume covering ellipsoid estimation in kernel-defined feature spaces," Proceeding of the 17th European Conference on Machine Learning

[12] G.R.G. Lanckriet, L.E. Ghaoui, C. Bhattacharyya and M.I. Jordan. "A robust minimax approach to classfication," J. Machine Learning Research, vol.3, pp. 555-582, 2002

[13] J. Platt,. "Fast training of support vector machines using sequential minimal optimization," Advances in kernel methods—Support vector learning, pp. 185–208, Cambridge, MIT Press, 1999

[14] X.D Zhang, "Matrix Analysis and Applications," Qinghua University Press, Sep. 2004

[15] J. T. Kwok, I. W. Tsang and J. M. Zurada. "A Class of Single-Class Minimax Probability Machines for Novelty Detection," IEEE Transactions on Neural Networks, vol. 18, No. 3, pp. 778-785, May 2007.

[16] D. Wang, D.S. Yeung, and E.C.C. Tsang. "Structured one-class classification," IEEE Trans. on Systems, Man, and Cybernetics - Part B: Cybernetics, vol. 36, No.6, pp. 1283-1294, 2006.

Appendix: Derivation (13) from (12), use Woodbury formula to the second term of (12) under the condition that $\lambda\mathbf{HKH}$ is taken as a whole.

$$\frac{\lambda}{n}\mathbf{KH}\left(\mathbf{I}+\frac{\lambda}{n}\mathbf{HKH}\right)^{-1}\mathbf{HK}=\frac{\lambda}{n}\mathbf{KH}\left(\frac{n}{\lambda}(\mathbf{HKH})^{-1}-\frac{n}{\lambda}(\mathbf{HKH})^{-1}\left(\mathbf{I}+(\frac{\lambda}{n}\mathbf{HKH})^{-1}\right)^{-1}\frac{n}{\lambda}(\mathbf{HKH})^{-1}\right)\mathbf{HK}$$

$$=\mathbf{K}-\mathbf{H}^{-1}\left(\mathbf{I}+(\frac{\lambda}{n}\mathbf{HKH})^{-1}\right)^{-1}\frac{n}{\lambda}\mathbf{H}^{-1}=\mathbf{K}-\left(\mathbf{H}\left(\mathbf{I}+(\frac{\lambda}{n}\mathbf{HKH})^{-1}\right)\left(\frac{\lambda}{n}\mathbf{H}\right)\right)^{-1}$$

$$=\mathbf{K}-\left(\frac{\lambda}{n}\mathbf{H}+\mathbf{K}^{-1}\right)^{-1}$$

Put the above result back to (12) and the derivation is proved.