Structured Learning from Data for Novelty Detection by Linear Programming

Aimin Feng, XuejunLiu, Bin Chen

***Abstract*—Novelty detection involves modeling the normal patterns for detecting any divergence from this behavior. Our recently proposed algorithm, Glabal&Local One Class Classifier (GLocal OCC), can solve this problem by maximizing the margin between the hyperplane and the origin through embedding the global information into the OCSVM framework. In this paper, we propose Linear Programming (LP) GLocal OCC (lpGLocal OCC) instead of the Quadratic Programming optimization to speed up GLocal OCC. By minimizing the average functional distance of the overall samples to the hyperplane, the lpGLocal OCC can attract the optimal hyperplane towards the centre of the data without using the origin anymore. Borrow off-the-shelf LP solver, this novel algorithm can be implemented easily and process solve large datasets rapidly. Results on benchmark datasets show that lpGLocal OCC not only has the comparable generalization power compared with the GLocal OCC besides its efficiency, but also has better generalization than (lp)OCSVM due to its structured learning approach.**

***Index Terms*—Linear Programming, Novelty Detection, Quadratic Programming, Structured Learning**

I. Introduction

Novelty detection [1] can be implemented by one class classification which usually differentiates the normal patterns from the outliers. These detection tasks can be found in many real-world scenarios like machine faulty diagnosis, network intrusion detection and document classification etc. Traditionally, novel patterns are detected by either estimating the probability density function of the normal patterns or by quantile estimation. However, these approaches both depend critically on the parametric form of the density function and can fail miserably when this assumption is incorrect.

Instead of estimating the density or quantile, a simpler task is to model the support of the data distribution directly. Through finding the boundary to enclose the normal patterns appropriately, this approach minimizes the domain of the normal patterns by geometric shapes such as hypersphere or hyperplane. The philosophy behind this derived from Vapnik who advocates that never solve a more general problem as the intermediate process. [2]

Manuscript received December 30, 2008. This work is supported by the National Nature Science foundation of China, No.60703016, No.60603029 and Jiangsu province Nature Science foundation No.BK2007589.

Aimin Feng is with the Computer Science & Technology College, Nanjing University of Aeronautics & Astronautics, 210016, Nanjing, P.R. China (phone No. is +86-25-84896490-12108, Fax:+86-25-84892848, e-mail:amfeng@nuaa.edu.cn).

XuejunLiu is with the Computer Science & Technology College, Nanjing University of Aeronautics & Astronautics (e-mail: xuejun.liu@nuaa.edu.cn)

Bin Chen is with the Dept. of Computer Science, Yangzhou University 225009, Yangzhou, P.R. China (e-mail:b.chen@nuaa.edu.cn).

As the typical algorithm of hypersphere model, Support vector data description (SVDD) [3] uses a small ball to enclose most of the data. Making use of the kernel trick, SVDD also works well on high-dimensional data by replacing the dot products between patterns with the corresponding kernel functions. In order to get much tighter domain, some ellipsoid algorithms are proposed to surround the overall data instead of using the hypersphere. Among them, Minimum Volume Enclosing Ellipsoid (MVEE) [4] uses the unit ellipsoid to cover most of the normal patterns, while the Mahalanobis Ellipsoidal Learning Machine (MELM) [5] optimizes the ellipsoid by using the M-metric radius. As for the Minimum Volume Covering Ellipsoid (MVCE) [6], it optimizes the covariance matrix and the ellipsoid's radius simultaneously.

Beside balls, hyperplane is usually employed to detect the outlier. As the state-of-the-art SVM applied on one class classification, One-Class SVM (OCSVM)[7] finds the optimal hyperplane by separating the normal patterns from the origin with maximal margin. Further analyses the OCSVM, we have found this algorithm has the local learning property since it neglects the whole data's distribution due to the use of the Euclidian metric margin. In contrast, Single-Class Minimax Probability Machine (SCMPM) [8], another approach based on hyperplane model, seeks the smallest half space for normal patterns through the global learning way by using mean and covariance of the data. While in fact, global and local learning from the data are both very important for the classifier design.

Motivated by unifying the global and local information into an integrated framework, we recently proposed a Global & Local One Class Classifier (GLocal OCC) [9] through embedding the distribution issues into OCSVM framework. In this way, the GLocal OCC not only incorporates the global and local learning into a unified classifier, but also provides a general way to extend the classical SVM algorithms for considering the global issues of the data. Here we call the global and local learning as *structured learning* briefly. Moreover, the optimization of the GLocal OCC is also Quadratic Programming (QP) alike the OCSVM so that can be solved by the standard SVM implementation such as the SMO optimization.

In this paper, we propose the Linear Programming (LP) GLocal OCC (lpGLocal hereafter) for further reducing the computational cost of GLocal OCC. Through minimizing the mean functional distance of the whole samples to the hyperplane, the lpGLocal OCC can automatically attract the optimal hyperplane toward the centre of the data. In this way, the lpGLocal OCC possesses performs the following advantages compared with the above algorithms:

- Needn't compel the origin to act as the representative of the outliers anymore

- Easy to implement by the LP solver or other optimization approaches, such as interior-point method
- Computationally more efficient
- has more powerful generalization ability for structured learning

The rest of the paper is organized as following: Section 2 outlines our GLocal OCC algorithms, including its linear and kernel form. Section 3 presents the novel lpGlocal OCC for speeding up Glocal OCC. Section 4 shows the experimental results on benchmark datasets. Finally, some conclusions are drawn in Section 5.

II. GLOCAL ONE-CLASS-CLASSIFIER

In order to unify the global and local issues into an integrated framework, our GLocal OCC incorporates the covariance matrix into the original OCSVM for taking into account the normal patterns' distribution when maximizing its margin. In the following, the linear and kernel formulation will be displayed respectively.

A. Linear GLocal OCC

Given a set of normal patterns $\mathrm{X} = \{x_1, x_2, \cdots, x_n\}$, GLocal OCC tries to find the hyperplane by maximizing the margin and minimizing given data's scatter degree denote by covariance matrix in the objective function:

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

Where $\nu \in (0,1)$ is the parameter which characterizes the fraction of support vectors and outliers, named ν-property[7]. ξ_i is the slack variables used to penalize the normal samples lying on the negative half space. Both of them are the original parameters of the OCSVM. Here the $\boldsymbol{\Sigma}$ denotes the scatter-ness of the given samples which represent the global issue of the input data, while the other items are the same as those in OCSVM, which try to find support vectors referred to as the local manner of the normal patterns. The item λ is the regularized factor that regulates the balance between the new term $\mathbf{w}^{\mathbf{T}}\boldsymbol{\Sigma}\mathbf{w}$ and the original $\mathbf{w}^{\mathbf{T}}\mathbf{w}$ in the OCSVM. Here the value of λ is no less than zero and the bigger of this value, the more emphasize on the global issue of the data.

Transforming the primal style into the corresponding dual formulation:

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

Where $\alpha = \left[\alpha_1, \cdots, \alpha_n\right]^{\mathbf{T}}$, $\mathbf{1} = \left[1, \cdots, 1\right]^{\mathbf{T}}$.

Notice that the above dual does not work in the input space where defined by the inner product $\mathbf{X}^{\mathbf{T}}\mathbf{X}$ as the kernel trick usually does, but is replaced by $\mathbf{X}^{\mathbf{T}}\left(\mathbf{I}+\lambda\boldsymbol{\Sigma}\right)^{-1}\mathbf{X}$ which maps the samples into a new feature space for finding the optimal hyperplane. When this hyperplane is mapped back into the input space, it becomes a nonlinear boundary which undoubtedly has more separable ability than the linear hyperplane.

From the dual form(2), we know GLocal OCC is also a QP problem, which means the solving process is almost the same as the OCSVM. That is, off-the-shelf QP solver or some decomposition methods such as SMO [10] can be exploited even without much modification.

Given an unseen data, the decision function is described as:

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

From the above primal(1), the dual(2) and the decision function(3), we notice if the factor λ is set to zero, GLocal OCC will reduce to the original OCSVM. We therefore conclude that GLocal OCC is the extended form of OCSVM by incorporating more consideration of the global information.

A. Kernel GLocal OCC

For utilizing the kernel trick in the dual form, all the terms of $\mathbf{X}^{\mathbf{T}}\left(\mathbf{I}+\boldsymbol{\lambda}\boldsymbol{\Sigma}\right)^{-1}\mathbf{X}$ are denoted by the inner product. So the $\boldsymbol{\Sigma}$ is described as:

$$\boldsymbol{\Sigma} = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{X}\mathbf{X}^{\mathbf{T}} - \frac{\mathbf{1}}{\mathbf{n}^{\mathbf{2}}}\mathbf{X}\mathbf{1}\mathbf{1}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}} = \frac{\mathbf{1}}{\mathbf{n}}\mathbf{X}\mathbf{H}\mathbf{X}^{\mathbf{T}} \tag{4}$$

Here $\mathbf{H} = (\mathbf{I} - \frac{\mathbf{1}}{\mathbf{n}}\mathbf{1}\mathbf{1}^{\mathbf{T}})$, where $\mathbf{I}$ is identity matrix. So by using the following Woodbury formula [11]:

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

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

$$\left(\mathbf{I}+\lambda\boldsymbol{\Sigma}\right)^{-1} = \mathbf{I} - \frac{\lambda}{n}\mathbf{X}\mathbf{H}(\mathbf{I} + \frac{\lambda}{n}\mathbf{H}\mathbf{X}^{\mathbf{T}}\mathbf{X}\mathbf{H})^{-1}\mathbf{H}\mathbf{X}^{\mathbf{T}} \tag{5}$$

By adopting the kernel trick, Equation (2) then becomes [12]:

$$\begin{aligned} &\min \ \frac{1}{2}\alpha^{T}(\mathbf{K} - \frac{\lambda}{n}\mathbf{K}\mathbf{H}(\mathbf{I} + \frac{\lambda}{n}\mathbf{H}\mathbf{K}\mathbf{H})^{-1}\mathbf{H}\mathbf{K})\alpha \\ &s.t. \ \ \alpha^{T}\mathbf{1} = 1, \quad 0 \leq \alpha \leq \frac{1}{\nu n}\mathbf{1} \end{aligned} \tag{6}$$

where $\mathbf{K} = \mathbf{X}^{\mathbf{T}}\mathbf{X}$ is the kernel matrix. This is also a standard QP. Moreover, when $\mathbf{K}$ is invertible, by using the Woodbury formula, (6) can be further simplified:

$$\begin{aligned} &\min \ \frac{1}{2}\alpha^{T}(\mathbf{K}^{-1} + \frac{\lambda}{n}\mathbf{H})^{-1}\alpha \\ &s.t \ \ \alpha^{T}\mathbf{1} = 1, \quad 0 \leq \alpha \leq \frac{1}{\nu n}\mathbf{1} \end{aligned} \tag{7}$$

The decision function (3) is also changed to:

$$\mathbf{F}(\mathbf{x}) = \mathbf{Sgn}\left[\boldsymbol{\alpha}^{\mathbf{T}}(\tilde{\mathbf{K}} - \frac{\lambda}{n}\mathbf{K}\mathbf{H}(\mathbf{I} + \frac{\lambda}{n}\mathbf{H}\mathbf{K}\mathbf{H})^{-1}\mathbf{H}\tilde{\mathbf{K}} - \rho\mathbf{1}\right] \tag{8}$$

Where $\tilde{\mathbf{K}}$ represents the kernel matrix between normal

patterns and testing data points. Since it is not a square matrix, Equation (3) can not be further simplified as(7). Here $\mathbf{F(x)}$ and $\mathbf{Sgn}(\bullet)$ are the vector representation of $f(\boldsymbol{x})$ and $\mathrm{sgn}(\bullet)$ in(3).

III. Linear Programming Glocal OCC

The above GLocal OCC is solved computational expensive. In order to further improve the efficiency of the computation, inspired by the LP version of OCSVM [13], we proposed the novel lpGLocal OCC to replace the QP solver of GLocal OCC. Instead of directly maximize the margin between the hyperplane and the origin, the lpGLocal OCC minimizes the output of the whole samples to the hyperplane. In this way, the lpGLocal OCC not only avoids the drawback of arbitrary taking origin as the outlier, but also attracts the optimal hyperplane located on the place of the minimum positive half space by adopting structured learning which is the spirit of GLocal OCC.

A. Linear lpGLocal OCC

For the hard margin case, the target function finds a hyperplane in input space to separate the normal patterns from the abnormals. This hyperplane is pulled onto the input samples with the restriction that each data point should always be in the positive half space. By minimizing the mean value of the distance from each pattern to the hyperplane, the objective function can be achieved as following:

$$\begin{aligned} &W(\boldsymbol{\alpha},\boldsymbol{\rho}) = \frac{1}{n}\left[\left(\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\left(\mathbf{I}+\lambda\boldsymbol{\Sigma}\right)^{-1}\mathbf{X}-\rho\mathbf{1}\right)\right]^{T}\mathbf{1} \\ &s.t.\ \ \alpha^{T}\mathbf{1}=1,\quad \alpha\geq 0 \\ &\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\left(\mathbf{I}+\lambda\boldsymbol{\Sigma}\right)^{-1}\mathbf{X}\geq\rho\mathbf{1} \end{aligned} \tag{9}$$

In the(9), the added constraint implies that all the given patterns should be in the positive half space.

For avoiding the bad effects of the noise and outliers, here we introduce the soft margin:

$$\begin{aligned} &\min W(\boldsymbol{\alpha},\boldsymbol{\rho}) \\ &=\frac{1}{n}\left[\left(\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\left(\mathbf{I}+\lambda\boldsymbol{\Sigma}\right)^{-1}\mathbf{X}-\rho\mathbf{1}\right)+\frac{1}{v}\boldsymbol{\xi}\right]^{T}\mathbf{1} \\ &s.t.\ \ \alpha^{T}\mathbf{1}=1,\quad 0\leq\alpha\leq\frac{1}{vn}\mathbf{1}, \\ &\boldsymbol{\alpha}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\left(\mathbf{I}+\lambda\boldsymbol{\Sigma}\right)^{-1}\mathbf{X}\geq\rho\mathbf{1}-\boldsymbol{\xi},\boldsymbol{\xi}\geq 0 \end{aligned} \tag{10}$$

Where $\boldsymbol{\xi}=[\xi_{1},\xi_{2},\cdots,\xi_{n}]$ denotes the slack variables of all the given patterns.

Since there is no any change applied to the decision function, we still use(3) to decide the unseen data belongings.

B. Kernel form

Similar to the kernel trick usually done, the linear lpGLocal OCC can work on high-dimensional data if the $\left(\mathbf{I}+\lambda\boldsymbol{\Sigma}\right)^{-1}$ is denoted by the dot products between the given data. Since this work has been fulfilled during the kernelized of GLocal OCC with(5) , we can use this result directly and derive the following nonlinear case:

$$\begin{aligned} &\min W(\boldsymbol{\alpha},\rho) \\ &=\frac{1}{n}\left[(\boldsymbol{\alpha}^{\mathbf{T}}(\mathbf{K}-\frac{\lambda}{n}\mathbf{K}\mathbf{H}(\mathbf{I}+\frac{\lambda}{n}\mathbf{H}\mathbf{K}\mathbf{H})^{-1}\mathbf{H}\mathbf{K})\right]^{T}\mathbf{1} \\ &-\rho+\frac{1}{v}\boldsymbol{\xi}^{T}\mathbf{1} \\ &s.t.\ \ \alpha^{T}\mathbf{1}=1,\quad 0\leq\alpha\leq\frac{1}{vn}\mathbf{1}, \\ &\boldsymbol{\alpha}^{\mathbf{T}}(\mathbf{K}-\frac{\lambda}{n}\mathbf{K}\mathbf{H}(\mathbf{I}+\frac{\lambda}{n}\mathbf{H}\mathbf{K}\mathbf{H})^{-1}\mathbf{H}\mathbf{K})\geq\rho\mathbf{1}-\boldsymbol{\xi} \\ &\boldsymbol{\xi}\geq 0 \end{aligned} \tag{11}$$

If $\mathbf{K}$ is invertible, the above formula can be further simplified:

$$\begin{aligned} &\min_{\alpha}\ \left[\alpha^{T}(\mathbf{K}^{-1}+\frac{\lambda}{n}\mathbf{H})^{-1}-\rho\mathbf{1}+\frac{1}{v}\boldsymbol{\xi}\right]^{T}\mathbf{1} \\ &s.t.\ \ \alpha^{T}\mathbf{1}=1,\quad 0\leq\alpha\leq\frac{1}{vn}\mathbf{1}, \\ &\boldsymbol{\alpha}^{\mathbf{T}}(\mathbf{K}^{-1}+\frac{\lambda}{n}\mathbf{H})^{-1}\geq\boldsymbol{\rho}-\boldsymbol{\xi},\ \boldsymbol{\xi}\geq 0 \end{aligned} \tag{12}$$

The decision function is the same as(8).

C. Time complexity analysis of lpGLocal OCC

Since lpGLocal OCC is the LP improvement of the original QP solving, here we omit the computation consume of the matrix in the (lp)GLocal OCC and only care about the computational complexity of LP and QP. Using the simplex method or interior-point method, the computational complexity of LP is approximately $o(n)$, while that of the standard SVM QP solvers (MINOS, CPLEX, LOQO, MATLAB QP routines) is $O(n^3)$. Therefore, the optimization advantage of LP algorithms is obvious.

IV. Experiment

A. Criteria of evaluation

In the following experiment, we mainly use the error rate to evaluate the generalization of the algorithms.

- False Negative(FN): rate of the normal patterns are misclassified as outliers, also called the first error ;
- False Positive (FP): rate of the outliers are misclassified as normal patterns. Also called the second error;
- Balance Loss (BL): the mean of the above errors, BL=$(\mathrm{FP}+\mathrm{FN})/2$

Obviously, the lower of the above criteria, the better performance of the algorithm.

B. Effect of the regularized factor λ

In order to investigate the effect of the regularized factor λ on the classifier, we perform experiments on a toy problem with the normal data come from a banana-shaped set. 50 normal points are used for training the lpGLocal OCC with a RBF kernel:

TABLE I. THE FN/FP/BL RESULTS WITH DIFFERENT AMOUNTS OF COVARIANCE INFORMATON ON BANANA-SHAPED DATASET($\nu = 0.1$)

λ	0	10	100	1000	10000
FN	0.1660	0.1705	0.1470	0.1375	0.0750
FP	0.2445	0.2350	0.2095	0.1545	0.2820
BL	0.2053	0.2027	0.1782	0.1460	0.1785

TABLE II. FP/ FN /BL RESULTS ON THE UCI DATA OF (LP)OCSVM AND (LP)GLOCAL OCC

FP/ FN /BL / Data Sets （Tr:TeT:TeO,D）	Unstructured learning		Structured learning	
	OCSVM	lpOCSVM	GLocal OCC	lpGLocal OCC
Biomed (102:25:67,5)	0.1418/0.1520/0.1469	0.2881 /0.1240/0.2060	***0.1313*/0.1120/*0.1217***	0.1403/0.1240/0.1321
Breast Cancer (367:9:241,9)	0.0228/0.0604/0.0416	0.0344 /0.0901/0.0623	**0.0274/0.0473*0.0373***	0.0257/0.1055/0.0656
Heart (123:41:139,13)	0.5424/0.2781/0.4103	0.4856 /0.4063/0.4459	***0.5165*/0.2531/*0.3848***	0.595/0.2000/0.3975
Import (71:17:71,25)	0.177/0.3353/0.2564	0.1282/0.3765/0.2523	**0.1817/*0.2294*/*0.2056***	0.2141/0.2294/0.2217
Ionosphere (80:45:126,34)	0.0317/0.1556/0.0937	0.0246 /0.2289/0.1267	0.0325/0.1156/0.0740	**0.0349/0.1067/0.0708**
Sonar (89:22:9,60)	0.1165/0.6591/0.3878	0.0299 /0.5545/0.2922	0.1351/0.5409/0.3380	**0.0959/0.4545/0.2752**
Arrhythmia (90:47:183,278)	0.4612/0.1191/0.2902	0.4475/0.1425/0.2950	0.3798/0.1596/0.2697	***0.3923/0.1404/0.2664***

$$K(x, y) = e^{-\|x-y\|^2/\sigma} \tag{13}$$

Here we set σ the mean distance between pair-wise points. For testing, we use another 200 normal points and 200 outliers outside the banana-shaped region. Table I shows the results on FP/ FN / BL (averaging over 10 repetitions) when different amounts of covariance information is used.

From the results of Table I, we can see that the generalization ability of classifier gets improved with more distribution information obtained by enlarging λ . Of course, if this information is overestimated, the performance will slow the increasing trend or even worse than the original ones.

C. Results on benchmark datasets

Here we list the performances of (lp)GLocal OCC and (lp)OCSVM on seven binary classes from the UCI machine learning repository on the first column of Table II. These results are listed by dimension from low to high, the （Tr:TeT:TeO,D） means the numbers of Training data, the Testing normal patterns and the Testing Outliers, D means the dimension of the datasets. Here we still follow the steps in [14] to take the larger class as normal data and the other as outliers, and then randomly sample 80% of the normal patterns for training, the remaining 20% of the normal patterns and all the outliers for testing.

Here we also use the RBF kernel of the(13). In accordance with the experiment setup we have reported on GLocal OCC, here we still use the grid search for finding the optimal kernel parameter σ and the regularized factor λ in (lp)GLocal OCC. Set $\nu = 0.1$ in all algorithms.

Since the LP algorithms are obviously superior to QP in computational complexity, here we omit the runtime but only compare the performance of the 4 algorithms divided into two groups according to their learning way. To reduce statistical variability, average results of 10 repetitions are reported in Table II. The italic and bold font denotes the best result of each data set according to the Balance Loss. From analyses, we can conclude the following results:

- Comparing the structured learning algorithms (lp)GLocal OCC with unstructured learning (lp)OCSVM respectively,, we notice that BL of (lp)GLocal OCC is better than its original algorithm in all seven datasets except the comparable result of lpGLocal OCC and lpOCSVM on Breast Cancer dataset. These results sufficiently prove that considering both the global and local information is more reasonable than only considering the local information as (lp)OCSVM does.
- Further analyses the reason for the better performance of the (lp)GLocal OCC, we found the small values of BL obtained are mainly ascribed to the lower values of FN compared with (lp)OCSVM. It is reasonable since (lp)GLocal OCC considers the target data's distribution when finding their decision boundary. We also notice that (lp)GLocal OCC possibly leads to large FP since its enlarged boundary has the risk to include the space of the outliers. However, this increasing of FP is usually slower than the decreasing of FN, so we can get the improved results of BL. This further proves that it is reasonable to take into account data's distribution for unstructured learning algorithms.
- Comparing the results of two structured learning algorithms, we can see the performance of lpGLocal OCC is comparable to GLocal OCC since it works better than GLocal OCC on three dataset denoted by bold and italic character. Particularly, for the Sonar dataset, LP algorithm performs better than the QP solver even by 6 percents.

V. Conclusion

In this paper, we proposed a linear programming approach lpGLocal OCC for accelerating our former model GLocal OCC. Through minimizing the average functional distance of the whole samples to the hyperplane, the lpGLocal OCC can attract the optimal hyperplane towards the centre of the data distribution. So the lpGLocal OCC need not to repel the hyperplane away from any arbitrary point outside the data distribution as the GLocal OCC does. As the result of the linear programming, this novel algorithm is easy to implement and able to process the large datasets rapidly. Experimental results on benchmark have shown that the lpGLocal OCC has comparable generalization power compared with the GLocal OCC besides its computation efficiency. In future work, inspired by sTructure OCC (TOCC) [15] which further considers the data distribution in delicate granularity, we will extend lp(GLocal) OCC to work on finer clusters within the normal patterns.

Acknowledgment

The authors thank the constructive discussion with Professor S. Chen of the Parnec group.

References

[1] Markou, M. and S. Singh, *Novelty detection: a review – part 1: statistical approaches.* Signal Processing, 2003. **83**(12): p. 2481-2497.

[2] Vapnik, V., *Statistical Learning Theory*. 1998, New York: Addison-Wiley

[3] Tax, D. and R.P. Duin, *Support Vector Data Description.* Machine Learning, 2004. **54**(1): p. 45-66.

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

[5] Wei, X.K., G.B. Huang, and Y.H. Li. *Mahalanobis Ellipsoidal Learning Machine for One Class Classification*. in *Proceedings of the Sixth International Conference on Machine Learning and Cybernetics*. 2007. Hong Kong: Washiington DC,USA:IEEE Computer Society 3528-3533.

[6] Dolia, A., C. Harris, J. Shawe-Taylor, et al., *Kernel ellipsoidal Trimming.* Computational Statistics and Data Analysis, 2007. **52**(1): p. 309-324.

[7] Schölkopf, B., J.C. Platt, and J. Shawe-Taylor, *Estimating the support of a high-dimensional distribution.* Neural Computation, 2001. **13(**7): p. 1443-1471.

[8] Lanckriet, G.R.G., L.E. Ghaoui, C. Bhattacharyya, et al., *A robust minimax approach to classfication.* J. Machine Learning Research, 2002. **3**: p. 555-582.

[9] Feng, A., B. Chen, and X. Liu. *Learning the Boundary of One Class Classifier Globally and Locally*. in *IEEE International Conference on Cybernetics and Intelligent Systems*. 2008. Chengdu, China: IEEE.364-369

[10] Platt, J. *Fast training of support vector machines using sequential minimal optimization*. in *Advances in kernel methods-Support vector learning*. 1999. Cambridge: MIT Press.185-208

[11] Platt, J. Fast training of support vector machines using sequential minimal optimization. in Advances in kernel methods-Support vector learning. 1999. Cambridge: MIT Press.185-208.

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

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

[14] James, T.K., I.W. Tsang, and J.M. Zurada, *A Class of Single-Class Minimax Probability Machines for Novelty Detection.* IEEE Transactions on Neural Networks, 2007. **18**(3): p. 778-785.

[15] Wang, D., D.S. Yeung, and E.C.C. Tsang, *Structured one-class classification.* IEEE Trans. on Systems, Man, and Cybernetics - Part B: Cybernetics, 2006. **36**(6): p. 1283-1294.