سیزدهمین کنفرانس
آمار ایران
۲-۴ شهریور ۱۳۹۵
13th Iranian Statistics Conference
Shahid Bahonar University of Kerman
23-25 August 2016

انجمن آمار ایران
IRANIAN STATISTICAL SOCIETY 1990

دانشگاه شهید باهنر کرمان

تاریخ: ۱۳۹۵/۶/۴

شماره: ۳۱۰/۲۴۸/آی

بسم الله الرحمن الرحیم

گواهی ارائه مقاله

بدینوسیله گواهی می‌شود

سرکار خانم ماندانا محمدی

مقاله خود را با عنوان:

An Outlier Pruning Preprocessing Approach for Support Vector Machine

در سیزدهمین کنفرانس آمار ایران ارائه نموده‌اند.

ضمن ارج نهادن به حضور ایشان امیدواریم که از این همایش علمی بهره‌مند شده باشند.

سایر همکاران: مجید سرمد

ماه بانو ماتا
دبیر علمی کنفرانس

محسن مددی
دبیر کنفرانس

http://isc13.uk.ac.ir
Email: ISC13@Conf.uk.ac.ir
Tel: +98 3431322260 Fax: +98 3433257157

دبیر خانه: کرمان، دانشگاه شهید باهنر کرمان
دانشکده ریاضی و کامپیوتر، بخش آمار

سیزدهمین کنفرانس
آمار ایران
13th Iranian Statistics Conference
Shahid Bahonar University of Kerman
23-25 August 2016

An Outlier Pruning Preprocessing Approach for Support Vector Machine

Mandana Mohammadi*, Majid Sarmad,

Department of Statistics, Ferdowsi University of Mashhad,
P. O. Box 1159, Mashhad 91775, Iran

Abstract

Support Vector Machine which is invented by Vapnik and Cortes in 1995, belongs to the statistical learning theory. Its application, has been tremendously increased over the years due to its prominent theoretical properties. The basic idea behind the support vector machine is to find the optimal hyper plane for linearly separable data. However the patterns that are not linearly separable can be transformed from original space into new space by means of the famous kernel function such as linear, polynomial, rbf and etc.. The presence of outlying observation can adversely affect the performance of support vector machine and will lead to the subsidence of its accuracy. In this paper, we provide a graphical depiction of data by using the high breakdown robust measure, namely the Mahalanobis distance based on the re–weighted minimum covariance determinant estimator. The so called method, "outlier map" is very popular in the multivariate robust statistics. It can be used to depict the structure of the data with any dimension. Using data from both simulation and real world studies, illustrated that the outlier map based on the robust Mahalanobis distance is the ability to recognise the outlying and misclassified samples in the data.

Keywords: Support Vector Machine, Outliers, Robust statistics, Mahalanobis Distance, re–weighted Minimum Covariance Determinant estimator
Mathematics Subject Classification (2010): MSC[2010] 62H30; 68T05; 62P10.

1 Introduction

In last few years there has been intense interest in Support Vector Machine (SVM), due to its high generalization capability used in many applications of data mining, engineering, and bioinformatics. SVM owes its reputation to its high classification accuracy, easiness of geometrically interpretation and very strong fundamental theory. Although, SVM try to find among different hyperplane candidate by maximization the margin, but in the presence of misclassified instance, it can not be able to find the optimal solution. In order to rectify this problem, soft margin that allows instances to fall within the margin and even on the wrong side of the separating hyperplane, has been invented (Cortes and Vapnik (1995)). The misclassified instances and the points inside the margin will be penalized through variable ξ_i.

Nevertheless, very few real world practical problems are linearly separable and thus no hyperplane found to separate the examples of two classes. To solve this issue, one can map the data from original space to a higher dimensional space namely feature space, and find separating hyperplane in the new space.

*Speaker: Manda.mohammadi@stu.um.ac.ir

Where the mapping function can be defined as $\Phi()$. So, the other prominent property of SVM for the non-linearity separable data set is its compatibility with kernel function ($K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi^T(\mathbf{x}_i)\phi(\mathbf{x}_j)$). Thus, the training of an SVM consists of the following minimization problem:

$$\min_{\mathbf{w},\xi} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \tag{1}$$

subject to,

$$y_i(w.\mathbf{x}_i - b) \geq 1 - \xi_i \quad 1 \leq i \leq n \tag{2}$$

Where w represents the hyperplane normal vector and b is the bias term. ξs are slack variables which allow for penalized constraint violation. C is the parameter controlling the trade-off between a large margin and less constrained violation.

However, the training data may polluted by the existence of some atypical observation and affect SVM models. In many applications, outlier detection as well as outlier removal is a first step in the statistical analysis process. One conventional way of recognising multivariate outliers in a multivariate normal dataset is the computation of Mahalanobis distance. The estimates of the location and scatter are utilized in the calculation of this distance to identify the far observations form the bulk of the data. Due to extremely sensitivity of the classical estimators to the presence of outliers in the data, the robust estimators of the multivariate location vector and scatter matrix should be used.

To remedy the aforementioned problem, based the idea of outlier map of Debruyne (2009), we used a robust version of Mahalanobis distance in the support vector machine problem. The outlier map is one of the common visual methods for the outliers identification in multivariate robust statistics [Rousseeuw and Van Zomeren (1990), Hubert and Engelen (2004) and Hubert, Rousseeuw and Vanden Branden (2005)]. We provide a graphical depiction of data which takes advantage of the robust Mahalanobis distance based on Minimum Covariance Determinant estimator (MCD). In Section 4, the application of outlier map on simulated and real data is considered.

The organization of the paper is as follows. Next section, contains description of the MCD and re–weighted version of this estimator. The algorithmic schemes of robust SVM classifier is described in 3. Detailed structure of simulated and real data is given in 4. In this Section, we apply our methods to a simple but illustrative toy-example and real data, then the experiment results are presented. Finally, we conclude the paper in 5.

2 The Minimum Covariance Determinant estimator

One of the first affine equivariant and highly robust estimators of multivariate location and scatter is the minimum covariance determinant estimator of Rousseeuw (1985). The property of resistivity to the outlying points, makes the MCD as a very handy tool in the robust statistics community.

Given a dataset with n p-dimensional observations, The MCD's goal is to find subset of size h (where $\frac{n}{2} < h < n$) out of n which has the smallest covariance determinant.

$$H_0 = \underset{H}{\operatorname{argmin}} \quad det(cov(x_i | i \in H)) \tag{3}$$

To illustration propose, the artificial dataset generated by Hawkins, Bradu, and Kass (Hawkins,Bradu and Kass (1984)) has been considered. The data set consists of 75 observations. The first 14 observations are outliers, created in two groups: 1-10 and 11-14. The classical methods can only detect the observations 12, 13 and 14, but by using MCD, this observation can be easily unmasked. A scatter plot of the data is shown in Figure 1, together with the classical and the robust 97.5% tolerance ellipse.

The classical tolerance ellipse is defined as the set of p-dimensional points x whose Mahalanobis distance

$$MD(x) = \sqrt{(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})} = \sqrt{\chi^2_{p,0.975}}. \tag{4}$$

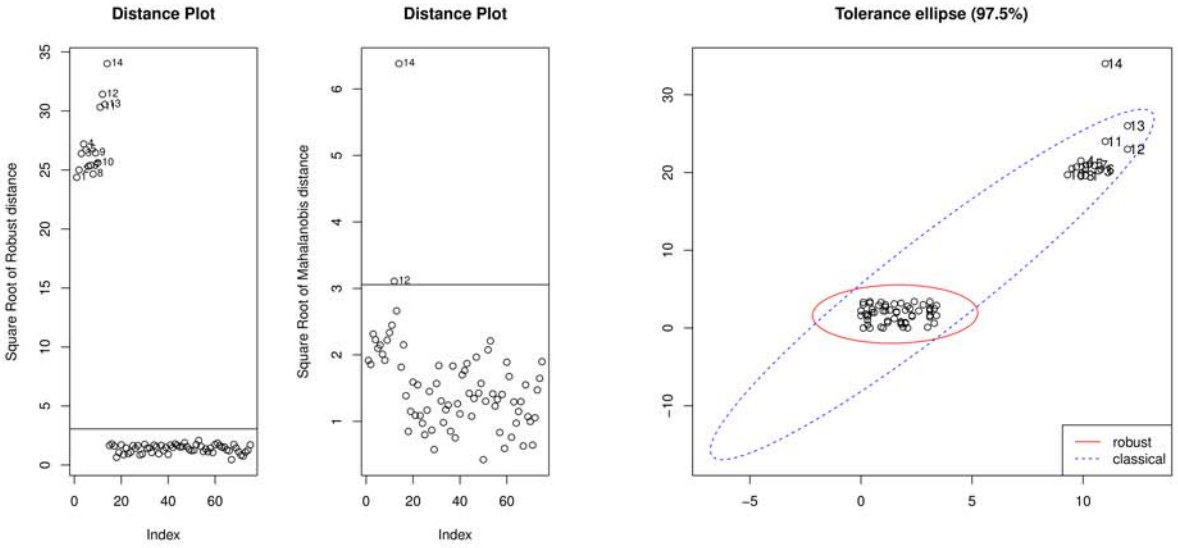

Figure 1: Bivariate Hawkins, Bradu, and Kass (hbk) data with classical and robust Mahalanaobis distanse (left panel) and tolerance ellipse (right panel).

The Mahalanobis distance $MD(\mathbf{x}_i)$ should tell us how far away x_i is from the center of the cloud, relative to the size of the cloud. Here is $\bar{\mathbf{x}}$ the sample mean and $\mathbf{S}$ the sample covariance matrix. It can be seen from Figure 1, this tolerance ellipse tries to cover all observations. As a result, only one observation can be consider as as mild outliers. On the other hand, the robust tolerance ellipse based on the MCD, is much smaller and does not include the outlying points. As can be seen from Figure 1, the outlying points are putting aside and consequently, the standards estimates of location and covariance can be computed based on the pure subset of observations, h.

The MCD estimator of location $\bar{\mathbf{X}}_{MCD} = \frac{1}{h} \sum_{i \in H0} x_i$, is the average of the aforementioned subset and MCD estimator of the scale is the covariance matrix of the subset times a factor which is the multiplication of the consistency and finite sample correction. According to Lopuhaa and Rousseeuw (1991) when $\lfloor h = \frac{n+p+1}{2} \rfloor$, the MCD has its highest breakdown point.

The MCD estimator can be computed in a reasonable time using the FAST–MCD algorithm of Rousseeuw and Driessen (1999); however, they are statistically not efficient. In order to carry robustness and efficiency, the re–weighted procedure is proposed. This procedure consists of omitting the observations with Mahalanobis distances which exceed a certain cut off value.

The re–weighted version of MCD estimation of mean vector is

$$\bar{\mathbf{X}}_{RMCD} = \frac{\sum_{i=1}^{n} w_i \mathbf{x}_i}{\sum_{i=1}^{n} w_i}, \tag{5}$$

and covariance matrix

$$\mathbf{S}_{RMCD} = c_{\eta,p} * d_{\gamma,\eta}^{n,p} \frac{\sum_{i=1}^{n} w_i (\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD})(\mathbf{x}_i - \bar{\mathbf{x}}_{RMCD})'}{\sum_{i=1}^{n} w_i} \tag{6}$$

where $c_{\eta,p}$ is used to make $\mathbf{S}_{RMCD}$ consistent under the multivariate normal distribution and $d_{\gamma,\eta}^{n,p}$ is the finite sample correction factor (see Croux and Haesbroeck (1999) and Pison, Van Aelst and Willems (2002))).

The weights are defined as indicator function based on the robust Mahalanobis distance as follows

$$w_i = \begin{cases} 1 & RD_i \leq q_\eta \\ 0 & \text{Otherwise} \end{cases} \tag{7}$$

where

$$RD_i = \sqrt{\sum_{i=1}^{n} (\mathbf{x}_i - \bar{\mathbf{x}}_{MCD})' \mathbf{S}_{MCD}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}_{MCD})} \tag{8}$$

and q_η ($\eta = 0.975$ (Rousseeuw and Driessen (1999))) is the quantile of the chi-square distribution with p degrees of freedom. To ease of readability, the re–weighted version of MCD abbreviated as RMCD.

3 Robust SVM classifier

In this section, the construction of robust SVM through the so–called robust method is presented. As mentioned in the previous sections, a criteria for outlier recognition is robust Mahalanobis distance based on the RMCD. The computation of robust SVM is feasible in the kernel space according to Debruyne (Debruyne (2009)). The so called SVM based on the RMCD, abbreviated as RMCD–SVM and can be computed as follows. Consider the problem of binary classification, the training data are given as

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n), \quad \text{where } \mathbf{x}_i \in \mathbb{R}^p \quad \text{and} \quad y_i \in \{-1, +1\}, \tag{9}$$

and let K be a kernel function. In the binary classification problem, the negative label class can be shown by $y = -1$ and the positive label class by the $y = +1$. The observations in the positive class is presented by n_+ and negative class by n_-.
The observation with large value of robust Mahalanobis distance must be omitted. The clean portion of the data can be classified by conducting the traditional SVM. The robust SVM algorithm outline for the RMCD–SVM is described as follows;

Outlier pruning: For the observation of the positive and negative class, compute one of the robust outlyingness measure. Retain only the 50% of observations with lowest amount of outlyingness.

Training: Once the outliers are pruned,the traditional SVM is used to classify the remained portion of the data which is illustrated by $T = T_- \cup T_+$.

Model selection: Optimize the kernel parameter γ and SVM penalty parameter C, over the range $\{10^{-6}, 10^{-1}\}$ and $\{10^{-1}, 10\}$ respectively, by 10 fold cross validation. Grid search is used for optimization of the parameters.

The scatter plot of the RMCD Mahalanobis distance versus the amount classifying function is plotted. The vertical line, which crossing from zero, is drawn. Plot the samples of two classes by different symbol (positive class is specified by the circles and negative class by the cross).

4 Experiment and Analysis

In this section some artificial and numerical data have been used to check the performance of the outlier map. Due to visualization problem, the case of a two-dimensional input space, i.e. $\mathbf{x}_i \in \mathbb{R}^2$ is considered.

4.1 Experiment with simulated dataset

To verify the performance of RMCD–SVM, 50 observation is generated from bivariate normal distribution for each class, with $\mu_1 = 0$, $\mu_2 = -3.5$ and standard deviation equal to one, where μ_1 and μ_2 are the mean of positive and negative class, respectively. Different simulation scheme has been considered, to check the performance of outlier map.

The clean data, in which all of the observation are from the same distribution and there is not any atypical point in the data. Contaminated data, refer to existence of some outlier in the data, for example, the distribution of 20% of data are different form the original data. The remaining scheme devoted to the misclassified and mixture of contaminated and misclassified data. The structure of simulation is given in the Table 1.

Table 1: Different scenario for simulation study

		Class I			Class II		
Case	**Tot. No. Obs.**	Label	Mean	Total	Label	Mean	Total
		I	μ_I	50	-	-	-
Clean	100	-	-	-	II	μ_{II}	50
		I	μ_I	45	I	μ_{II}	3
Misclassified	100	II	μ_I	5	II	μ_{II}	47
		I	μ_I	45	II	μ_I^*	3
Contaminated	100	I	μ_I^*	5	II	μ_{II}	47
		I	μ_I	42	I	μ_{II}	3
Misclassified–Contaminated	100	I	μ_I^*	4	II	μ_{II}	42
		II	μ_I	4	II	μ_{II}^*	5

The first part of the results devoted to the RMCD-SVM algorithm, using the artificial dataset. Two dimensional graph of the binary SVM classification is illustrated with the so called RMCD outlier map in the Figure 2–5.

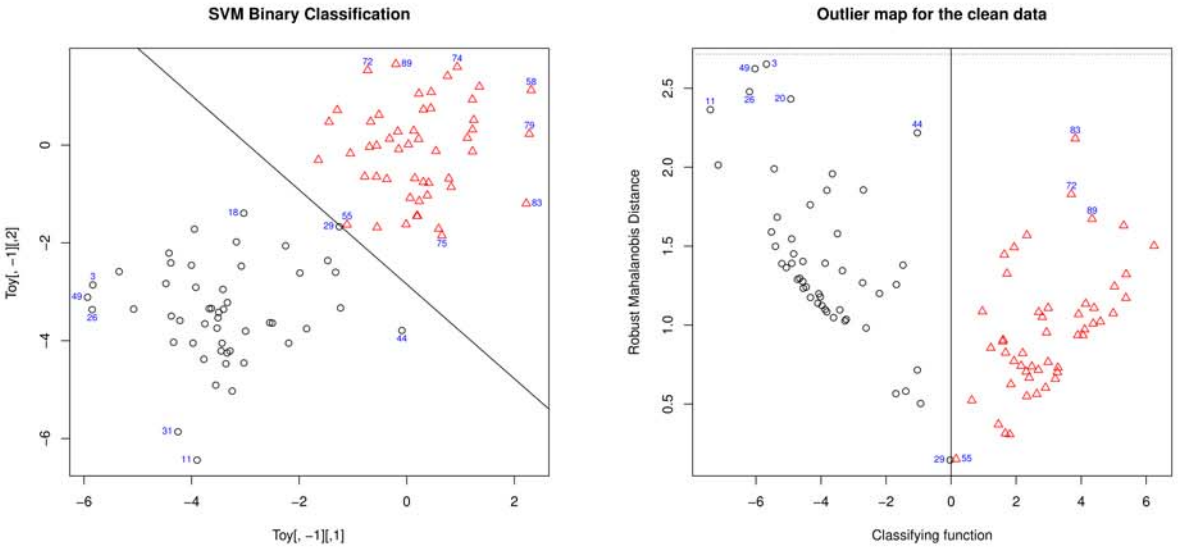

Figure 2: The plot of SVM binary classification (left panel) and the Mahalanobis distance based on the the RMCD versus the value of the classifier for the clean (right panel) data. The dotted line is the is the 97.5% quantile of the chi-square distribution with p degrees of freedom.

As can be seen from Figure 2, the homogeneity of the clean data is shown in right panel of Figure 2,

which show neither the misclassified nor the outliers. There are two risky data points, which are classified correctly, but they are very near to the boundary line, as can be seen from RMCD outlier map as well. The observations which have been placed further form the center of the data, are shown in both of the graphs.

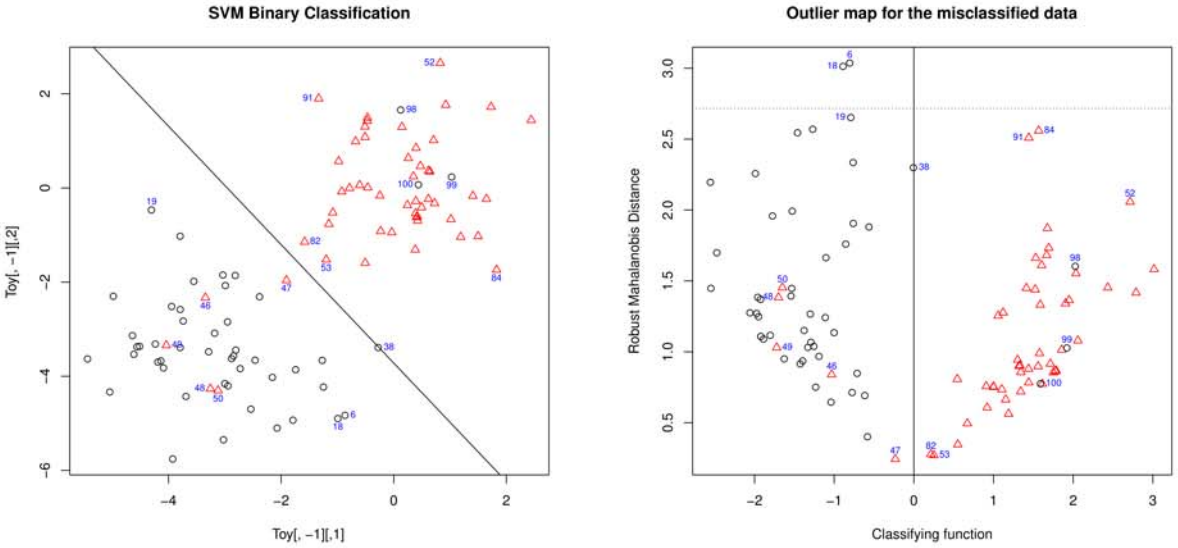

Figure 3: The plot of SVM binary classification (left panel) and the Mahalanobis distance based on the the RMCD versus the value of the classifier for the misclassified (right panel) data. The dotted line is the is the 97.5% quantile of the chi-square distribution with p degrees of freedom.

In Figure 3, the misclassified problem is presented. Likewise the previous example, the problematic observations are tagged in each class. The RMCD outlier map can detect respectively 5 and 3 misclassified observations form negative and positive class. The number of detected misclassified observations exactly coincide the number of generated misclassified data. Moreover observations 6 and 18 are detected as outliers. As is clear from the left panel of Figure 3, the mentioned observations are at the larger distance from the centre. Three observation are located very near to the the boundary line and observation 38 is placed on the boundary line, which also can be seen in the outlier map of Figure 3. Figure 4 depict the presence of outlying points in the data. This observations are clearly far from the bulk of the data, stated in the down left and up right corner of the left panel of Figure 4. The mentioned points can be detected by the outlier map. Observation 44 and 3 are very near to the boundary, which is presented by outlier map as well. The observation 8 from the negative class and 89 and 94 from the second positive class are further from the center of the data, that is showed by their large amount of robust Mahalanobis distance. From Figure 5, it appears that, the outlier map using robust Mahalanobis distance, has the ability of detection of a more sever case, which contained the misclassified and outlying points simultaneously.

4.2 Experiment with Real Dataset

The performance of the suggested techniques is assessed through a real dataset. The famous Iris (two class only) data has been used. The Iris data set contains 3 classes of 50 instances each (Setosa, Versicolour and Virginica) , where each class refers to a type of iris plant. In this paper, we only consider two classes whose labels are Versicolour and Virginica. The data set contains measurements of four variables - sepal length and width and petal length and width.

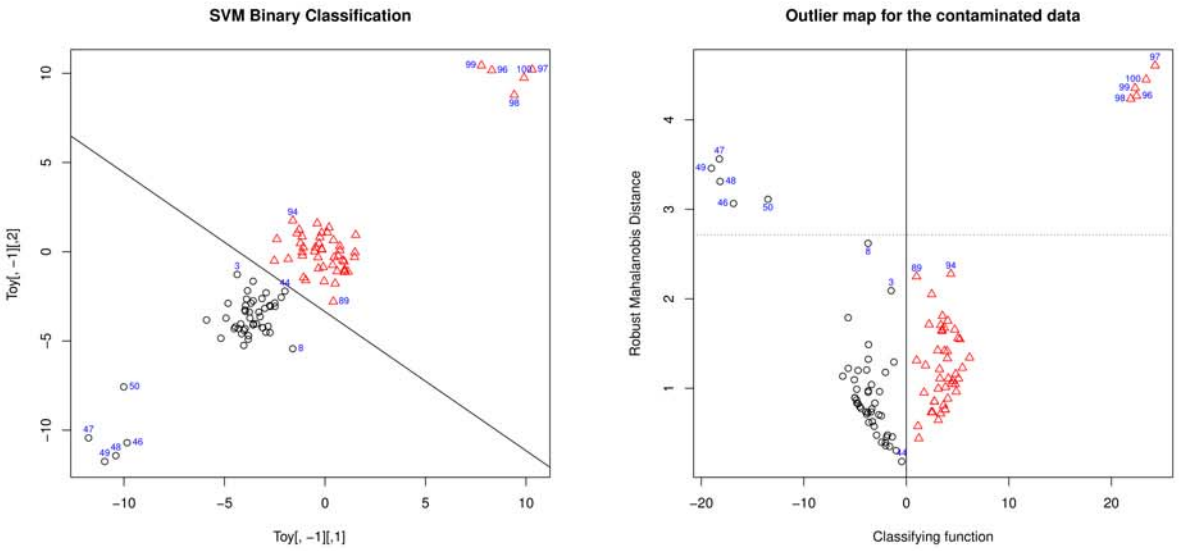

Figure 4: The plot of SVM binary classification (left panel) and the Mahalanobis distance based on the the RMCD versus the value of the classifier for the contaminated (right panel) data. The dotted line is the is the 97.5% quantile of the chi-square distribution with p degrees of freedom.

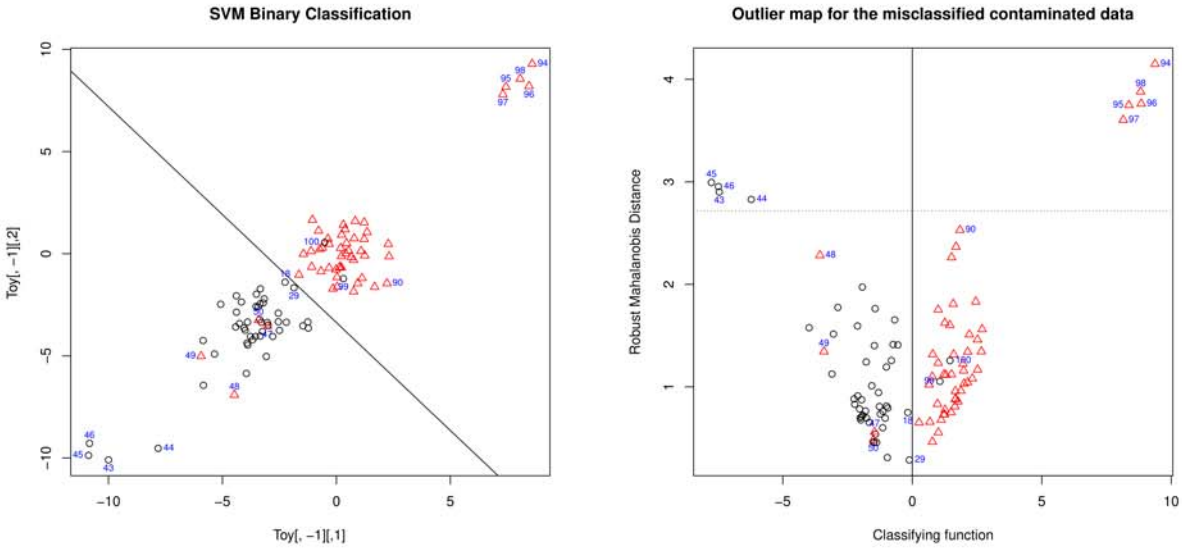

Figure 5: The plot of SVM binary classification (left panel) and the Mahalanobis distance based on the the RMCD versus the value of the classifier for the misclassified–contaminated (right panel) data. The dotted line is the is the 97.5% quantile of the chi-square distribution with p degrees of freedom.

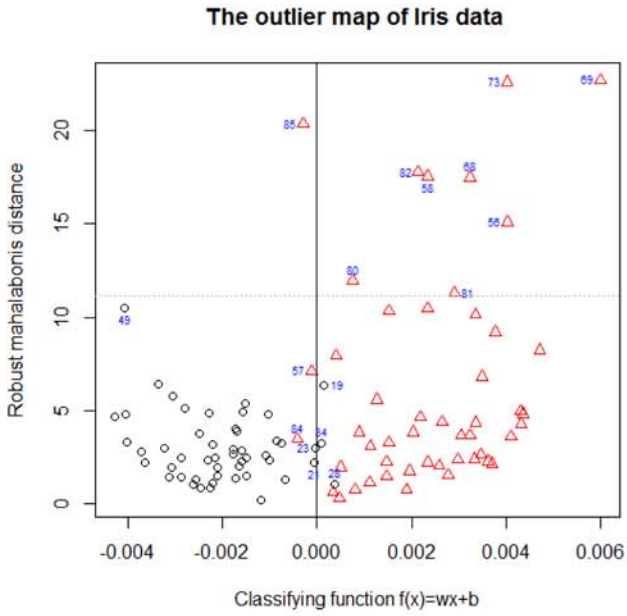

Figure 6: The plot of the Mahalanobis distance based on the RMCD versus the value of the classifier for the two class Iris dataset. The dotted line is the is the 97.5% quantile of the chi-square distribution with p degrees of freedom.

From the Figure 6, there is one outliers which is the misclassified observation as well in the Versicolour class and 8 outliers in the Virginica class.

5 Conclusion

The presence of outlier can adversely affect the performance of support vector machine. In order to rectify this problem robust methods have been utilized. In this paper a new robust SVM algorithm has been introduced. The high breakdown robust methods namely Mahalanobis distance based on the re–weighted minimum covariance determinant is used in this paper. The performance of the proposed robust SVM algorithm has been assessed by artificial and real data set. With respect to the Table 1 and the outlier map of the simulated data in Figure 2–5, it can be seen that the number of unusual observations, whether being the outlying point or misclassified observation, exactly coincide with their generated scenario.

References

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory* (pp. 144-152). ACM.

Bottou, L., Cortes, C., Denker, J. S., Drucker, H., Guyon, I., Jackel, L. D., ... and Vapnik, V. (1994, October). Comparison of classifier methods: a case study in handwritten digit recognition. *In International Conference on Pattern Recognition* (pp. 77-77). IEEE Computer Society Press.

Cortes, C., and Vapnik, V. (1995). Support-vector networks. *Machine learning*, 20(3), 273-297.

Croux, C., and Haesbroeck, G. (1999). Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. *Journal of Multivariate Analysis*, 71(2), 161-190.

Debruyne, M. (2009). An outlier map for support vector machine classification. *The Annals of Applied Statistics*, 1566-1580.

Duda, R. O., Hart, P. E., and Stork, D. G. (2012). *Pattern classification*. John Wiley & Sons.

Friedman, J. (1996). Another approach to polychotomous classifcation. *Dept. Statist., Stanford Univ., Stanford, CA, USA, Tech. Rep.*

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1-3), 389-422.

Hastie, T., and Tibshirani, R. (1998). Classification by pairwise coupling. *The annals of statistics*, 26(2), 451-471.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Hawkins, D. M., Bradu, D., and Kass, G. V. (1984). Location of several outliers in multiple-regression data using elemental sets. *Technometrics*, 26(3), 197-208.

Hubert, M., and Engelen, S. (2004). Robust PCA and classification in biosciences. *Bioinformatics*, 20(11), 1728-1736.

Hubert, M., Rousseeuw, P. J., and Vanden Branden, K. (2005). ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47(1), 64-79.

Karatzoglou, A., Smola, A., Hornik, K., and Zeileis, A. (2004). kernlab-an S4 package for kernel methods in R.

Knerr, S., Personnaz, L., and Dreyfus, G. (1990). Single-layer learning revisited: a stepwise procedure for building and training a neural network. In *Neurocomputing* (pp. 41-50). Springer Berlin Heidelberg.

Lopuhaa, H. P., and Rousseeuw, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices.*The Annals of Statistics,* 229-248.

Pison, G., Van Aelst, S., and Willems, G. (2002). Small sample corrections for LTS and MCD.*Metrika*, 55(1-2), 111-123.

Rousseeuw, P. J. (1985). Multivariate estimation with high breakdown point. *Mathematical statistics and applications*, 8, 283-297.

Rousseeuw, P. J., and Van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association*, 85(411), 633-639.

Rousseeuw, P. J., and Driessen, K. V. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3), 212-223.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer Science & Business Media.