15

Mahalanobis Support Vector Machines Made Fast and Robust

Xunkai Wei†‡, Yinghong Li‡, Dong Liu†, Liguang Zhan†
†*Beijing Aeronautical Engineering Technology Research Center*
‡*Air Force Engineering University*
CHINA

1. Introduction

As is known to us, common Euclidean distance based SVMs are easily influenced by outliers in given samples and might subsequently cause big prediction errors in testing processes. Therefore, many scholars propose various preprocessing methods such as whitening or normalizing the data to a sphere shape to remove the outliers and then call the routine SVM methods to build a more reasonable machine. However, since Euclidean distance is often sub-optimal especially in high dimension learning problem and might cause the learning machine fail due to the ill-conditioned Gram kernel, then it is necessary to find some more efficient and robust way to resolve the problem, which is the motivation of this chapter.

The Mahalanobis distance is superior to Euclidean Distance in handling with outliers and is widely used in statistics and machine learning area. Currently there are some methods in building SVMs combined with Mahalanobis distances. Some of them use it in the kernel and replace common kernel by a Mahalanobis one in SVMs. Some of them use it in preprocessing phase to remove the outlier first and then build SVMs using common methods. Others use it in the postprocessing phase to extract key support vectors for speedup and efficiency. Most of them achieve superior performances compared with SVM counterpart. However, it should be pointed out that the complexity of the combined algorithm is the most concerned factor in building such an algorithm.

As is known to us, none of them incorporates the Mahalanobis distance into models, which tradeoff the complexity and performance in the same algorithm meantime and make the algorithm more robust. The obvious feature of this new method is that there is no more necessary to remove the outlier first, since it is already considered and will be identified automatically in the model. It is also expected to improve and simplify the whole learning process efficiently.

One Class Classification (OCC) (Scholkopf, 2001) now becomes an active topic in machine learning domain. One Class Support Vector Machines (OCSVM) is firstly proposed via constructing a hyperplane in kernel feature space which separates the mapped patterns from the origin with maximum margin. Support vector domain description (SVDD) (Tax, 1999) is another popular OCC method, which seeks the minimum hypersphere that encloses all the data of the target class in a feature space. In this way, it finds the descriptive area that

covers the data and excludes the superfluous space that results in false alarms existed in OCSVM.
However, although OCSVM does provide good representation for the classes of interest, it overlooks the discrimination issue between them. Moreover, the hypersphere model of SVDD is not flexible enough to give a tight description of the target class generally.
Therefore, in our previous works, we proposed two Mahalanobis distance based learning machine called QP-MELM and QP-MHLM respectively via solving their duals. However, as is suggested in (Löfberg, 2004), if both the primal form and dual form of an optimization problem are solvable, then the primal form is more commendable for approximation ability. Therefore, (Wei, 2007A) rewrote the MELM as a Second Order Cone Programming (SOCP) representable form and proposed a SOCP-MELM for class description. Applications to real world UCI benchmark datasets show promising results.
Recently, Wei etc al proposed a novel learning concept called enclosing machine learning (Wei, 2007D), which imitates the human being's cognition process, i.e. cognizing things of the same kind (To obtain a minimum bounding boundary for class description) and recognizing unknown things via point detection. Wei illustrated the concept using minimum volume enclosing ellipsoid learner for one class description and extended it to imbalanced data set classification. Except this, (Wang, 2005) and (Liu, 2006) proposed two SVDD based pattern classification algorithms (called SSPC and MEME respectively for simplicity) for imbalanced data set, which can also been classified to enclosing machine leaning's framework .
This chapter will be organized as follows. First, review of Mahalanobis distance\property and related learning methods will be briefed. Then, the new optimization models based on linear programming for Data Description, Classification incorporating Mahalanobis distance will be proposed. Third, benchmark datasets experiments for classification and regression will be investigated in detail. Finally, conclusions and discussions will be made.

2. The Mahalanobis Distance

2.1 Definitions

Let $\mathbf{X}$ be a $m \times N$ sample matrix containing N random observations $\mathbf{x}_i \in R^m, i = 1, 2, \cdots, N$. The sample mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ can be concisely expressed in terms of sample matrix.

$$\begin{cases} \boldsymbol{\mu} = \dfrac{1}{N}\mathbf{X}\mathbf{1} \\ \boldsymbol{\Sigma} = \dfrac{1}{N}\mathbf{X}\mathbf{X}^T - \dfrac{1}{N^2}\mathbf{X}\mathbf{1}\mathbf{1}^T\mathbf{X}^T \end{cases} \tag{1}$$

where $\mathbf{1}$ is a $N-$dimensional all one vector.
If the covariance matrix is singular, it is difficult to calculate the inverse of $\boldsymbol{\Sigma}$. Instead, we can use the pseudoinverse $\boldsymbol{\Sigma}^+$ to approximate as $\boldsymbol{\Sigma}^+ = \mathbf{P}\boldsymbol{\Sigma}^{-1}\mathbf{P}^T$ using inverse of nonzero

eigenvalues. This gives the minimum squared error approximation to the true solutions. It should be noted that pseudoinverse restricts inversion to the range of the operator, i.e. the subspace where it is not degenerate. This is often unavoidable in high dimensional feature spaces. If the covariance is real symmetric and positive semidefinite, then the covariance matrix can be decomposed as $\mathbf{\Sigma} = \mathbf{P}^T\mathbf{G}\mathbf{P}$, and thus $\mathbf{\Sigma}^{-1} = \mathbf{P}^T\mathbf{G}^{-1}\mathbf{P}$. Then the Mahalanobis distance from a sample $\mathbf{x}$ to the population $\mathbf{X}$ is

$$d^2(\mathbf{x}, \mathbf{X}) = (\mathbf{x} - \boldsymbol{\mu})^T \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \tag{2}$$

2.2 The Mahalanobis Distance in Kernel Feature Space

In implicit high-dimensional feature space defined by kernel functions, the Mahalanobis distance can be represented in terms of the dot products of data maps. Suppose $\mathbf{X}^{\Phi}, \boldsymbol{\mu}^{\Phi}, \mathbf{\Sigma}^{\Phi}$ are the sample matrix, mean vector, and covariance matrix in the feature space, respectively. The centered kernel matrix is defined as

$$\mathbf{K}_C = \mathbf{K} - \frac{1}{N}\mathbf{E}\mathbf{K} - \frac{1}{N}\mathbf{K}\mathbf{E} + \frac{1}{N^2}\mathbf{E}\mathbf{K}\mathbf{E} \tag{3}$$

where $\mathbf{E}$ is a $N \times N$ all one matrix, $\mathbf{K} = \left\{ \Phi(\mathbf{x}_i) \bullet \Phi(\mathbf{x}_j) = k(\mathbf{x}_i, \mathbf{x}_j) \right\}_{i,j=1,\cdots,N}$ is a $N \times N$ symmetric matrix.

Using (1), we obtain

$$\mathbf{\Sigma}^{\Phi} = \mathbf{X}^{\Phi T}\mathbf{Z}^2\mathbf{X}^{\Phi} \tag{4}$$

where $\mathbf{Z} = (\frac{1}{N}(\mathbf{I} - \frac{1}{N}\mathbf{E}))^{\frac{1}{2}}$ is a $N \times N$ symmetric matrix, $\mathbf{I}$ is a $N \times N$ unit matrix.

The kernel Mahalanobis distance in the feature space can then be written as

$$\begin{aligned} d^2(\Phi(\mathbf{x}), \mathbf{X}^{\Phi}) &= (\Phi(\mathbf{x}) - \boldsymbol{\mu}^{\Phi})^T \mathbf{\Sigma}^{\Phi -1} (\Phi(\mathbf{x}) - \boldsymbol{\mu}^{\Phi}) \\ &\overset{k(\bullet)}{=} (k(\mathbf{X}, \mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N} k(\mathbf{X}, \mathbf{x}_i))^T \times (\mathbf{Z}\mathbf{M}^{-2}\mathbf{Z}) \\ &\times (k(\mathbf{X}, \mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N} k(\mathbf{X}, \mathbf{x}_i)) \end{aligned} \tag{5}$$

where $\mathbf{K} = k(\mathbf{X}, \mathbf{X}^T)$, $\mathbf{x}_i$ is the i th sample of $\mathbf{X}$, $\mathbf{M} = \mathbf{ZKZ}$ is symmetric and semidefinite matrix and thus $\mathbf{M}^{-2}$ can be calculated via singular value decomposition, i.e. $\mathbf{M} = \mathbf{U}\Lambda\mathbf{U}^T, \mathbf{M}^{-2} = \mathbf{U}\Lambda^{-2}\mathbf{U}^T$.

Using singular value decomposition method, we can easily conclude following theorem:

Theorem 1: Let the eigenstructures of the centered matrix $\mathbf{K}_C$ be $\mathbf{K}_C = \mathbf{Q}^T\mathbf{\Omega}\mathbf{Q}$, then the covariance matrix $\mathbf{\Sigma}^\Phi$ can be diagonalized as follows:

$$\mathbf{\Sigma}^\Phi = (\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{X}^{\Phi T})^T(\frac{1}{N}\mathbf{\Omega})(\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{X}^{\Phi T}) \tag{6}$$

where N is the number of samples.

Proof: Recall that the covariance matrix $\mathbf{\Sigma}^\Phi$ in the feature space, and suppose $\mathbf{\Sigma}^\Phi$ could be decomposed via Singular Value Decomposition (SVD): $\mathbf{\Sigma}^\Phi = \mathbf{P}^T\mathbf{G}\mathbf{P}$, we have

$$\begin{aligned}\mathbf{\Sigma}^\Phi &= \frac{1}{N}\sum_{i=1}^{N}(\Phi(\mathbf{x}_i) - \boldsymbol{\mu}^\Phi)(\Phi(\mathbf{x}_i) - \boldsymbol{\mu}^\Phi)^T \\ &= \frac{1}{N}\mathbf{X}^\Phi\mathbf{X}^{\Phi T} - \frac{1}{N^2}\mathbf{X}^{\Phi T}\mathbf{1}\mathbf{1}^T\mathbf{X}^\Phi = \mathbf{P}^T\mathbf{G}\mathbf{P}\end{aligned} \tag{7}$$

Notice that the eigenvectors necessarily lie in the span of the centered data, thus $\mathbf{P}$ can be written as following linear combination

$$\mathbf{P} = \boldsymbol{\theta}(\mathbf{X}^{\Phi T} - \frac{1}{N}\mathbf{E}\mathbf{X}^{\Phi T}) \tag{8}$$

where $\boldsymbol{\theta}$ is the coeffient matrix to be determined, $\mathbf{E}$ is a $N \times N$ all one matrix.

Multiplying (8) by $\mathbf{X}^{\Phi T} - \frac{1}{N}\mathbf{E}\mathbf{X}^{\Phi T}$ from the left side, and by $\mathbf{P}^T$ from the right side, we have (after substituting $\mathbf{X}^{\Phi T}\mathbf{X}^\Phi = \mathbf{K}$)

$$\frac{1}{N}(\mathbf{K}_C)^2\boldsymbol{\theta}^T = \mathbf{K}_C\boldsymbol{\theta}^T\mathbf{G} \tag{9}$$

Multiplying (9) by pseudoinverse $\mathbf{K}_C^+$ from the left side, we have

$$\frac{1}{N}\mathbf{K}_C\boldsymbol{\theta}^T = \boldsymbol{\theta}^T\mathbf{G} \tag{10}$$

Since matrix $\mathbf{G}$ is diagonal, and the centered kernel matrix can be decomposed as

$$\mathbf{K}_C = \mathbf{Q}^T\boldsymbol{\Omega}\mathbf{Q} \tag{11}$$

We can obtain

$$\boldsymbol{\theta} = \mathbf{DQ} \tag{12}$$

$$\mathbf{G} = \frac{1}{N}\boldsymbol{\Omega} \tag{13}$$

where $\mathbf{D}$ is some diagonal matrix.
Since the eigenvectors of the covariance matrix are orthogonal, we have

$$\begin{aligned}\mathbf{I} &= \mathbf{PP}^T \\ &= \boldsymbol{\theta}(\mathbf{X}^{\Phi T} - \frac{1}{N}\mathbf{E}\mathbf{X}^{\Phi T})(\mathbf{X}^{\Phi} - \frac{1}{N}\mathbf{X}^{\Phi}\mathbf{E})\boldsymbol{\theta}^T \\ &= \mathbf{DQK}_C\mathbf{Q}^T\mathbf{D} \\ &= \mathbf{D}\boldsymbol{\Omega}\mathbf{D}\end{aligned} \tag{14}$$

Therefore

$$\mathbf{D} = \boldsymbol{\Omega}^{-\frac{1}{2}} \tag{15}$$

Using the fact that

$$\begin{aligned}&\mathbf{K}_C = \mathbf{Q}^T\boldsymbol{\Omega}\mathbf{Q} \\ &\Rightarrow \mathbf{QK}_C = \boldsymbol{\Omega}\mathbf{Q} = \boldsymbol{\Omega}^{-1}\mathbf{QK}_C\mathbf{E} = \mathbf{QE} \\ &\Rightarrow \boldsymbol{\Omega}^{-1}\mathbf{Q}(\mathbf{K} - \frac{1}{N}\mathbf{EK} - \frac{1}{N}\mathbf{KE} + \frac{1}{N^2}\mathbf{EKE})\mathbf{E} = \mathbf{QE} \\ &\Rightarrow \boldsymbol{\Omega}^{-1}\mathbf{Q}(\mathbf{KE} - \frac{1}{N}\mathbf{EKE} - \mathbf{KE} + \frac{1}{N}\mathbf{EKE}) = \mathbf{QE} \\ &\Rightarrow \boldsymbol{\Omega}^{-1}\mathbf{Q0} = \mathbf{QE} = 0\end{aligned} \tag{16}$$

Thus from (8), using (12), (15) and (16), we have

$$\mathbf{P} = \mathbf{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{X}^{\Phi T} \tag{17}$$

Consequently, from (7), using (13), (17), we obtain

$$\mathbf{\Sigma}^{\Phi} = (\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{X}^{\Phi T})^{T}(\frac{1}{N}\mathbf{\Omega})(\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{X}^{\Phi T}) \tag{18}$$

And this ends the complete proof of Theorem 1.

According to Theorem 1, we can calculate the pseudoinverse $\mathbf{\Sigma}^{\Phi+}$ to approximate $\mathbf{\Sigma}^{\Phi-1}$ as

$$\mathbf{\Sigma}^{\Phi+} = N\mathbf{X}^{\Phi}\mathbf{Q}^{T}\mathbf{\Omega}^{-2}\mathbf{Q}\mathbf{X}^{\Phi T} \tag{19}$$

Thus (5) can be simplified as

$$\begin{aligned} d^{2}(\Phi(\mathbf{x}), \mathbf{X}^{\Phi}) = N(k(\mathbf{X}, \mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N} k(\mathbf{X}, \mathbf{x}_{i}))^{T} \\ \times(\mathbf{Q}^{T}\mathbf{\Omega}^{-2}\mathbf{Q}) \\ \times(k(\mathbf{X}, \mathbf{x}) - \frac{1}{N}\sum_{i=1}^{N} k(\mathbf{X}, \mathbf{x}_{i})) \end{aligned} \tag{20}$$

As we mentioned before, there often exists zero eigenvalues condition in high dimensional feature spaces, (7) is represented to approximate the true inverse of sample covariance matrix. Later, we will introduce another regularization method for avoiding the zero eigenvalues condition in Section 3.

3. Mahalanobis One Class SVMs

3.1 Mahalanobis One Class SVM

It is often beneficial to utilize the covariance matrix $\mathbf{\Sigma}$ and use the Mahalanobis distance instead. Given a set of unlabeled patterns $\{x_1, x_2, \cdots, x_N\}$, the OCSVM first maps them to the feature space $\mathbb{F}$ via a nonlinear map Φ. In the sequel, the OCSVM is obtained via solving

$$\min_{w,\xi_i \geq 0,\rho} \frac{1}{2}\mathbf{w}^T\mathbf{w} + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho$$
$$s.t.\begin{cases}\mathbf{w}^T x_i \geq \rho - \xi_i, \\ \xi_i \geq 0, i = 1,2,\cdots,N.\end{cases} \tag{21}$$

where $\mathbf{w}^T x = \rho$ is the decision hyperplane, $\xi_i \geq 0$ is slack variable, N is the number of samples. Using kernel trick, the corresponding dual

$$\max_{\alpha_i \geq 0} \frac{1}{2}\boldsymbol{\alpha}^T\mathbf{K}\boldsymbol{\alpha}$$
$$s.t.\begin{cases}\mathbf{0} \leq \boldsymbol{\alpha} \leq \frac{1}{\nu N}\mathbf{1} \\ \boldsymbol{\alpha}\mathbf{1} = 1.\end{cases} \tag{22}$$

is a quadratic convex optimization problem, where $\boldsymbol{\alpha}$ is the dual variable, $\mathbf{K} = \mathbf{X}^{\Phi T}\mathbf{X}^{\Phi}$ is a kernel matrix. By using the Mahalanobis distance metric instead of Euclidean distance metric, the primal now becomes:

$$\min_{w,\xi_i \geq 0,\rho} \frac{1}{2}\mathbf{w}^T\boldsymbol{\Sigma}^{-1}\mathbf{w} + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho$$
$$s.t.\begin{cases}\mathbf{w}^T\boldsymbol{\Sigma}^{-1} x_i \geq \rho - \xi_i, \\ \xi_i \geq 0, i = 1,2,\cdots,N.\end{cases} \tag{23}$$

where $\boldsymbol{\Sigma}$ is the sample covariance matrix.
Using $\mathbf{w} = \boldsymbol{\Sigma}\mathbf{u}$, then the separation hyperplane is now $\mathbf{u}^T x = \rho$ and the distance to the origin becomes $\frac{\rho}{\sqrt{\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}}}$. Therefore, (23) is equivalent to

$$\min_{\mathbf{u},\xi_i \geq 0,\rho} \frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u} + \frac{1}{\nu N}\sum_{i=1}^{N}\xi_i - \rho$$
$$s.t.\begin{cases}\mathbf{u}^T x_i \geq \rho - \xi_i, \\ \xi_i \geq 0, i = 1,2,\cdots,N.\end{cases} \tag{24}$$

The dual is as follows using optimality conditions:

$$\max_{\boldsymbol{\alpha}} \frac{1}{2}\boldsymbol{\alpha}^T \mathbf{X}^{\Phi T}\boldsymbol{\Sigma}^{-1}\mathbf{X}^{\Phi}\boldsymbol{\alpha}$$
$$s.t.\begin{cases}\mathbf{0} \le \boldsymbol{\alpha} \le \dfrac{1}{\nu N}\mathbf{1} \\ \boldsymbol{\alpha}^T\mathbf{1} = 1.\end{cases} \tag{25}$$

Note that the kernel trick is

$$\mathbf{X}^{\Phi T}\mathbf{X}^{\Phi} \overset{k(\cdot)}{=} \mathbf{K}, \tag{26}$$

Using the kernel trick and (21), the kernel Mahalanobis hyperplane learning machine can be written in kernel form as:

$$\max_{\boldsymbol{\alpha}} \frac{1}{2} N\boldsymbol{\alpha}^T \mathbf{K}\mathbf{Q}^T\boldsymbol{\Omega}^{-2}\mathbf{Q}\mathbf{K}\boldsymbol{\alpha}$$
$$s.t.\begin{cases}\mathbf{0} \le \boldsymbol{\alpha} \le \dfrac{1}{\nu N}\mathbf{1} \\ \boldsymbol{\alpha}^T\mathbf{1} = 1.\end{cases} \tag{27}$$

where the symbols are defined as previously noted.
We can easily conclude that (13) is a QP problem, thus can be efficiently solved via YALMIP[12].
As is mentioned before, there is uncertainty in the estimation of $\boldsymbol{\Sigma}$ in general. We can assume that $\boldsymbol{\Sigma}$ is only known to be within the set [5]

$$\{\boldsymbol{\Sigma} : \left\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^0\right\|_F \le r\} \tag{28}$$

Suppose $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^0 + r\Delta\boldsymbol{\Sigma}$, then according to the Cauchy-Schwarz inequality [14], we get
$\mathbf{u}^T\Delta\boldsymbol{\Sigma}\mathbf{u} \le \left\|\mathbf{u}\right\|_2 \left\|\Delta\boldsymbol{\Sigma}\mathbf{u}\right\|_2 \le \left\|\mathbf{u}\right\|_2 \left\|\Delta\boldsymbol{\Sigma}\right\|_F \left\|\mathbf{u}\right\|_2 = \mathbf{u}^T\mathbf{u}$
This holds of compatibility of the Frobenius matrix norm and the Euclidean vector norm and because $\left\|\Delta\boldsymbol{\Sigma}\right\|_F \le 1$. For $\Delta\boldsymbol{\Sigma}$ the unity matrix, this upper bound is attained.
Thus we have

$$\begin{aligned}\max_{\boldsymbol{\Sigma}:\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^0\right\|_F\le r}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}&=\mathbf{u}^T(\boldsymbol{\Sigma}^0+r\max_{\|\Delta\boldsymbol{\Sigma}\|_F\le 1}\Delta\boldsymbol{\Sigma})\mathbf{u}\\&=\mathbf{u}^T(\boldsymbol{\Sigma}^0+r\mathbf{I})\mathbf{u}\end{aligned}\tag{29}$$

where $r>0$ is fixed and $\|\bullet\|_F$ denotes the Frobenius norm, $\boldsymbol{\Sigma}^0$ is estimated via (1). Then the primal in (10) can be modified as

$$\begin{aligned}&\min_{\mathbf{u},\xi_i\ge 0,\rho}\max_{\boldsymbol{\Sigma}}\frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}\mathbf{u}+\frac{1}{\nu N}\sum_{i=1}^{N}\xi_i-\rho\\&s.t.\begin{cases}\mathbf{u}^T x_i\ge\rho-\xi_i,\\\left\|\boldsymbol{\Sigma}-\boldsymbol{\Sigma}^0\right\|\le r,\\\xi_i\ge 0, i=1,2,\cdots,N.\end{cases}\\&\Rightarrow\min_{\mathbf{u},\xi_i\ge 0,\rho}\frac{1}{2}\mathbf{u}^T\boldsymbol{\Sigma}_r\mathbf{u}+\frac{1}{\nu N}\sum_{i=1}^{N}\xi_i-\rho\\&s.t.\begin{cases}\mathbf{u}^T x_i\ge\rho-\xi_i,\\\xi_i\ge 0, i=1,2,\cdots,N.\end{cases}\end{aligned}\tag{30}$$

where $\boldsymbol{\Sigma}_r=\boldsymbol{\Sigma}^0+r\mathbf{I}$, now the sample covariance matrix is always nonsingular for $r>0$. Actually, this is a regularization method.
The dual is now

$$\begin{aligned}&\max_{\boldsymbol{\alpha}}\frac{1}{2}\boldsymbol{\alpha}^T\mathbf{X}^{\Phi T}\boldsymbol{\Sigma}_r^{-1}\mathbf{X}^{\Phi}\boldsymbol{\alpha}\\&s.t.\begin{cases}\mathbf{0}\le\boldsymbol{\alpha}\le\frac{1}{\nu N}\mathbf{1}\\\boldsymbol{\alpha}^T\mathbf{1}=1.\end{cases}\end{aligned}\tag{31}$$

By using the Woodbury formula [10]

$$(\mathbf{A}+\mathbf{BC})^{-1}=\mathbf{A}^{-1}-\mathbf{A}^{-1}\mathbf{B}(\mathbf{I}+\mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1}\tag{32}$$

and (4), we obtain

$$\begin{aligned}\boldsymbol{\Sigma}_r^{-1} &= (r\mathbf{I} + \mathbf{X}^{\Phi}\mathbf{Z}\mathbf{Z}\mathbf{X}^{\Phi T})^{-1} \\ &= \frac{1}{r}\mathbf{I} - \frac{1}{r}\mathbf{X}^{\Phi}\mathbf{Z}(r\mathbf{I} + \mathbf{Z}\mathbf{X}^{\Phi T}\mathbf{X}^{\Phi}\mathbf{Z})^{-1}\mathbf{Z}\mathbf{X}^{\Phi T}\end{aligned} \tag{33}$$

Using the kernel trick, (17) then becomes

$$\begin{aligned}&\max_{\boldsymbol{\alpha}} \frac{1}{2r}\boldsymbol{\alpha}^T(\mathbf{K} - \mathbf{K}\mathbf{Z}\mathbf{M}_r^{-1}\mathbf{Z}\mathbf{K})\boldsymbol{\alpha} \\ &s.t.\begin{cases}\mathbf{0} \le \boldsymbol{\alpha} \le \dfrac{1}{\nu N}\mathbf{1} \\ \boldsymbol{\alpha}^T\mathbf{1} = 1.\end{cases}\end{aligned} \tag{34}$$

where $\mathbf{M}_r = r\mathbf{I} + \mathbf{Z}\mathbf{K}\mathbf{Z}$. Again, we get a standard QP problem, and the inverse of the real symmetric and positive definite matrix $\mathbf{M}_r$ can be exactly estimated using stable and efficient eigenvalue decomposition method.

3.2 Mahalanobis Data Description Machine

Given a set of unlabeled patterns $\{x_1, x_2, \cdots, x_N\}$, the SVDD first maps them to the feature space $\mathbb{F}$ via a nonlinear map Φ. In the sequel, the SVDD is obtained via solving

$$\begin{aligned}&\min_{R,\xi_i \ge 0,\mathbf{c}} R^2 + C\sum_{i=1}^{N}\xi_i \\ &s.t.\begin{cases}d^2(\mathbf{c}, \mathbf{x}_i) \le R^2 + \xi_i, \\ \xi_i \ge 0, i = 1, 2, \cdots, N.\end{cases}\end{aligned} \tag{35}$$

where R is radius, $\xi_i \ge 0$ is slack variable, $\mathbf{c}$ is the center, $d(\bullet)$ is the given distance metric (the default is the Euclid norm) , C is a tradeoff that controls the size of sphere and the errors, N is the number of samples.
The corresponding dual

$$\max_{\alpha_i \geq 0} \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}_i) - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{x}_j)$$
$$s.t. \begin{cases} 0 \leq \alpha_i \leq C, i = 1, 2, \cdots, N \\ \sum_{i=1}^{N} \alpha_i = 1. \end{cases} \tag{36}$$

is a convex optimization problem, where $\alpha_i \geq 0$ is dual variable, $k(\bullet)$ is a kernel function that satisfies Mercer condition.

By using the Mahalanobis distance metric instead of Euclidean distance metric, the primal now becomes:

$$\min_{R, \xi_i \geq 0, \mathbf{c}} R^2 + C \sum_{i=1}^{N} \xi_i$$
$$s.t. \begin{cases} (\mathbf{x}_i - \mathbf{c})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{c}) \leq R^2 + \xi_i, \\ \xi_i \geq 0, i = 1, 2, \cdots, N. \end{cases} \tag{37}$$

where $\boldsymbol{\Sigma}$ is the sample covariance matrix.

The dual is as follows using optimality conditions:

$$\max_{\alpha_i \geq 0} \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \mathbf{x}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}_j$$
$$s.t. \begin{cases} 0 \leq \alpha_i \leq C, i = 1, 2, \cdots, N \\ \sum_{i=1}^{N} \alpha_i = 1. \end{cases} \tag{38}$$

Using the kernel trick, and following equations,

$$\mathbf{x}_i^T \mathbf{X} \overset{k(\bullet)}{=} k(\mathbf{x}_i, \mathbf{X}), \mathbf{c} = \sum_{i=1}^{N} \alpha_i \mathbf{x}_i \,, \tag{39}$$

the kernel Mahalanobis Ellipsoidal learning machine can be written in kernel form as:

$$\max_{\alpha_i \geq 0} \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{X}) \mathbf{Q}^T \boldsymbol{\Omega}^{-2} \mathbf{Q} k(\mathbf{X}, \mathbf{x}_i)^T$$
$$- \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j k(\mathbf{x}_i, \mathbf{X}) \mathbf{Q}^T \boldsymbol{\Omega}^{-2} \mathbf{Q} k(\mathbf{X}, \mathbf{x}_j) \tag{40}$$
$$s.t. \begin{cases} 0 \leq \alpha_i \leq C, i = 1, 2, \cdots, N \\ \sum_{i=1}^{N} \alpha_i = 1. \end{cases}$$

Accordingly we can obtain the following Mahanlanobis distance of the sample $\mathbf{x}$ from the center $\mathbf{c}$ in the feature space:

$$d^2(\mathbf{x}^\Phi, \mathbf{c}^\Phi) = N(\mathbf{x}^\Phi - \mathbf{c}^\Phi)^T \mathbf{X}^\Phi \mathbf{Q}^T \boldsymbol{\Omega}^{-2} \mathbf{Q} \mathbf{X}^{\Phi T} (\mathbf{x}^\Phi - \mathbf{c}^\Phi)$$
$$= N(k(\mathbf{X}, \mathbf{x}) - \sum_{i=1}^{N} \alpha_i k(\mathbf{X}, \mathbf{x}_i))^T \mathbf{Q}^T \boldsymbol{\Omega}^{-2} \mathbf{Q} \tag{41}$$
$$\times (k(\mathbf{X}, \mathbf{x}) - \sum_{i=1}^{N} \alpha_i k(\mathbf{X}, \mathbf{x}_i))$$

The parameters R, ξ_i can be determined by the following relations via KKT conditions:

$$\begin{cases} d^2(\mathbf{x}^\Phi, \mathbf{c}^\Phi) < R^2, \alpha_i = 0, \xi_i = 0 \\ d^2(\mathbf{x}^\Phi, \mathbf{c}^\Phi) = R^2, 0 < \alpha_i < C, \xi_i = 0 \\ d^2(\mathbf{x}^\Phi, \mathbf{c}^\Phi) = R^2 + \xi_i, \alpha_i = C, \xi_i > 0 \end{cases} \tag{42}$$

Again, we can assume that $\boldsymbol{\Sigma}$ is only known to be within the set

$$\{\boldsymbol{\Sigma} : \left\| \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^0 \right\|_F \leq r\} \tag{43}$$

where $r > 0$ is fixed and called regularization constant and $\left\| \bullet \right\|_F$ denote s the Frobenius norm, $\boldsymbol{\Sigma}^0$ is estimated via (1).

Then the primal in (37) can be modified as

$$\min_{R,\xi_i \geq 0,\mathbf{c}} \max_{\boldsymbol{\Sigma}} R^2 + C\sum_{i=1}^{N} \xi_i$$
$$s.t. \begin{cases} (\mathbf{x}_i - \mathbf{c})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \mathbf{c}) \leq R^2 + \xi_i, \\ \xi_i \geq 0, i = 1, 2, \cdots, N, \\ \left\| \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^0 \right\|_F \leq r \end{cases} \tag{44}$$

Suppose $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^0 + r\Delta\boldsymbol{\Sigma}$. Then for any given $\mathbf{v}$, according to the Cauchy-Schwarz inequality, we get

$$\mathbf{v}^T \Delta\boldsymbol{\Sigma}\mathbf{v} \leq \left\|\mathbf{v}\right\|_2 \left\|\Delta\boldsymbol{\Sigma}\mathbf{v}\right\|_2 \leq \left\|\mathbf{v}\right\|_2 \left\|\Delta\boldsymbol{\Sigma}\right\|_F \left\|\mathbf{v}\right\|_2 = \mathbf{v}^T \mathbf{v}$$

This holds of compatibility of the Frobenius matrix norm and the Euclidean vector norm and because $\left\|\Delta\boldsymbol{\Sigma}\right\|_F \leq 1$. For $\Delta\boldsymbol{\Sigma}$ the unity matrix, this upper bound is attained.

Thus for any given $\mathbf{v}$ and $\boldsymbol{\Sigma}^0$, we have

$$\begin{aligned} \min_{\boldsymbol{\Sigma}: \left\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^0\right\|_F \leq r} \mathbf{v}^T \boldsymbol{\Sigma}^{-1} \mathbf{v} &= \mathbf{v}^T (\boldsymbol{\Sigma}^0 + r \max_{\|\Delta\boldsymbol{\Sigma}\|_F \leq 1} \Delta\boldsymbol{\Sigma})^{-1} \mathbf{v} \\ &= \mathbf{v}^T (\boldsymbol{\Sigma}^0 + r\mathbf{I})^{-1} \mathbf{v} \end{aligned} \tag{45}$$

Therefore, (44) can be modified as

$$\min_{R,\xi_i \geq 0,\mathbf{c}} R^2 + C\sum_{i=1}^{N} \xi_i$$
$$s.t. \begin{cases} (\mathbf{x}_i - \mathbf{c})^T \boldsymbol{\Sigma}_r^{-1} (\mathbf{x}_i - \mathbf{c}) \leq R^2 + \xi_i, \\ \xi_i \geq 0, i = 1, 2, \cdots, N. \end{cases} \tag{46}$$

where $\boldsymbol{\Sigma}_r = \boldsymbol{\Sigma}^0 + r\mathbf{I}$, now the sample covariance matrix is always nonsingular for $r > 0$. Actually, this is a regularization method.
The dual is now

$$\max_{\alpha_i \geq 0} \sum_{i=1}^{N} \alpha_i \mathbf{x}_i^T \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_i - \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \mathbf{x}_i^T \boldsymbol{\Sigma}_r^{-1} \mathbf{x}_j$$

$$s.t. \begin{cases} 0 \leq \alpha_i \leq C, i = 1, 2, \cdots, N \\ \sum_{i=1}^{N} \alpha_i = 1. \end{cases} \tag{47}$$

By using the Woodbury formula

$$(\mathbf{A} + \mathbf{BC})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{B}(\mathbf{I} + \mathbf{CA}^{-1}\mathbf{B})^{-1}\mathbf{CA}^{-1} \tag{48}$$

we obtain

$$\begin{aligned} \boldsymbol{\Sigma}_r^{-1} &= (r\mathbf{I} + \mathbf{X}^{\Phi}\mathbf{ZZX}^{\Phi T})^{-1} \\ &= \frac{1}{r}\mathbf{I} - \frac{1}{r}\mathbf{X}^{\Phi}\mathbf{Z}(r\mathbf{I} + \mathbf{ZX}^{\Phi T}\mathbf{X}^{\Phi}\mathbf{Z})^{-1}\mathbf{ZX}^{\Phi T} \end{aligned} \tag{49}$$

Using the kernel trick, (47) then becomes

$$\begin{aligned} &\max_{\alpha_i \geq 0} \sum_{i=1}^{N} \alpha_i (k(\mathbf{x}_i, \mathbf{x}_i) - k(\mathbf{x}_i, \mathbf{X})\mathbf{Z}\mathbf{M}_r^{-1}\mathbf{Z}k(\mathbf{X}, \mathbf{x}_i)) \\ &- \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j (k(\mathbf{x}_i, \mathbf{x}_j) - k(\mathbf{x}_i, \mathbf{X})\mathbf{Z}\mathbf{M}_r^{-1}\mathbf{Z}k(\mathbf{X}, \mathbf{x}_j)) \end{aligned} \tag{50}$$

$$s.t. \begin{cases} 0 \leq \alpha_i \leq C, i = 1, 2, \cdots, N \\ \sum_{i=1}^{N} \alpha_i = 1. \end{cases}$$

where $\mathbf{M}_r = r\mathbf{I} + \mathbf{ZKZ}$. Again, the inverse of the real symmetric and positive definite matrix $\mathbf{M}_r$ can be estimated via singular value decomposition.

Note that above mentioned models are both QP based, which might converge slowly in large dataset case. In order to further reduce the model complexity of above Mahalanobis Distance based SVM, here we introduce a new linear programming based model for ellipsoidal data description.

Let $\{\Phi(\mathbf{x}_i), i = 1, \cdots, N\}$ be the images of the samples in feature space through mapping Φ. We first center all the samples in feature space, and then we can build ellipsoidal machine centered at origin enclosing a majority of the imaged vectors. Then according to (5), the distance from any sample $\mathbf{x}$ to the origin can be kernelized as

$$d^2 = \varphi(x_i)\Sigma^{-1}\varphi(x_i)^T = \varphi(x_i)\Sigma^{+}\varphi(x_i)^T = \left\|\sqrt{N}\mathbf{\Omega}^{-1}\mathbf{Q}\mathbf{k}_i\right\|_2^2 \tag{51}$$

where $\mathbf{k}_i := (k(\mathbf{x}_1, \mathbf{x}_i), \cdots, k(\mathbf{x}_N, \mathbf{x}_i))$.
Therefore, now we can rewrite the primal form as:

$$\begin{aligned} &\min_{R^2, \xi_i \geq 0} R^2 + C\sum_{i=1}^{N}\xi_i \\ &s.t. \begin{cases} \left\|\sqrt{N}\mathbf{\Omega}^{-1}\mathbf{Q}\mathbf{k}_i\right\|_2^2 \leq R^2 + \xi_i, \\ \xi_i \geq 0, i = 1, 2, \cdots, N. \end{cases} \end{aligned} \tag{52}$$

The Lagrange function for the primal form will be as follows:

$$L(R^2, \xi_i, \alpha_i, \beta_i) = R^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\beta_i\xi_i - \sum_{i=1}^{N}\alpha_i(R^2 - \left\|\sqrt{N}\mathbf{\Omega}^{-1}\mathbf{Q}\mathbf{k}_i\right\|_2^2 + \xi_i) \tag{53}$$

where $\alpha_i, \beta_i \geq 0$ are Lagrange multipliers or the dual variables. According to the KKT conditions, and equating the partial derivatives of L with respect to R^2, ξ_i to zero yields:

$$\frac{\partial L}{\partial R^2} = 1 - \sum_{i=1}^{N}\alpha_i = 0 \tag{54}$$

$$\frac{\partial L}{\partial \xi_i} = C - \beta_i - \alpha_i = 0 \tag{55}$$

From (55), we get

$$\beta_i = C - \alpha_i \geq 0 \tag{56}$$

From (56) and using $\alpha_i, \beta_i \geq 0$, we can obtain

$$0 \leq \alpha_i \leq C \tag{57}$$

Using (57), and substituting (54), (55) into (53) results in the dual problem:

$$\min_{\alpha} -\sum_{i=1}^{N} \alpha_i \left\| \sqrt{N}\boldsymbol{\Omega}^{-1}\mathbf{Q}\mathbf{k}_i \right\|_2^2$$
$$s.t. \begin{cases} \sum_{i=1}^{N} \alpha_i = 1 \\ 0 \le \alpha_i \le C \end{cases} \tag{58}$$

We can see that the optimization problem (58) is in a linear programming form. This LP form is superior to QP form in computational complexity.

From the solution of (58), the samples with $\alpha_i = 0$ will fall inside the ellipsoid. The samples with $\alpha_i > 0$ is called support vectors. Support vectors with $\alpha_i = C$ is called border support vectors. And the radius of the ellipsoid can be obtained using

$$R = \left\| \sqrt{N}\boldsymbol{\Omega}^{-1}\mathbf{Q}\mathbf{k}_{sv} \right\| \tag{59}$$

via using any border support vectors $\mathbf{x}_{sv}$.

Again, we assume that $\boldsymbol{\Sigma}$ is only known to be within the set $\{\boldsymbol{\Sigma} : \left\| \boldsymbol{\Sigma} - \boldsymbol{\Sigma}^0 \right\|_F \le r\}$,where $r > 0$ is fixed and called regularization constant and $\| \ \|_F$ denotes the Frobenius norm. Then, the kernelized distance formula from any sample $\mathbf{x}$ to the origin is as follows:

$$d^2 = \varphi(x_i)\Sigma_r^{-1}\varphi(x_i)^T = \varphi(x_i)\Sigma_r^{+}\varphi(x_i)^T = \left\| (\frac{1}{N}\boldsymbol{\Omega} + r\mathbf{I})^{-\frac{1}{2}}\boldsymbol{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{k}_i \right\|_2^2 \tag{59}$$

Therefore, the robust Mahalanobis data description in kernel form is

$$\min_{R^2, \xi_i \ge 0} R^2 + C\sum_{i=1}^{N} \xi_i$$
$$s.t. \begin{cases} \left\| (\frac{1}{N}\boldsymbol{\Omega} + r\mathbf{I})^{-\frac{1}{2}}\boldsymbol{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{k}_i \right\|_2 \le R^2 + \xi_i, \\ \xi_i \ge 0, i = 1, 2, \cdots, N. \end{cases} \tag{60}$$

where $\mathbf{k}_i := (k(x_1, x_i), k(x_2, x_i), \cdots, k(x_N, x_i))$.

And the dual form is

$$\min_{\alpha} -\sum_{i=1}^{N}\alpha_i \left\| (\frac{1}{N}\mathbf{\Omega}+r\mathbf{I})^{-\frac{1}{2}}\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{k}_i \right\|^2$$
$$s.t.\begin{cases}\sum_{i=1}^{N}\alpha_i = 1 \\ 0 \le \alpha_i \le C\end{cases} \tag{61}$$

Accordingly, we can get the ellipsoidal radius function in robust form for any sample $\mathbf{x}$,

$$f(\mathbf{x}) = \sqrt{N}\left\| (\frac{1}{N}\mathbf{\Omega}+r\mathbf{I})^{-\frac{1}{2}}\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{k}_{\mathbf{x}} \right\|_2 \tag{62}$$

4. Mahalanobis Classification SVMs

4.1 The main idea

Given a set of training data $\{(\mathbf{x}_1, y_1),(\mathbf{x}_2, y_2),\cdots,(\mathbf{x}_N, y_N)\}$, where $\mathbf{x}_i \in R^k, y_i \in \{1,-1\}$, $y_i = 1$ denotes target class, $y_i = -1$ denotes outlier class. Recall that SVM tries to find an optimum separation hyperplane by maximizing the margin as shown in Fig. 1(a). One can also try to find a hypersphere with a maximum separation shell which separates the target class from the outlier class shown in Fig.1 (b). Co-inspired by this idea, but instead of trying to find a sphere that provides a description of one class, for classification purposes, we would like to find a more compact and flexible hyper-ellipsoid that encloses all samples from one class (target class) but excludes all samples from the other class (outlier class). That is to say, we would like to find a hyper-ellipsoid $\mathrm{E}(R,\mathbf{c})$ that encloses all the target class samples and excludes all outlier class samples (see Fig.1 (c)). For the depicted example, we can see that our proposed MMEE method is tighter. Thus this can commendably reduce the risk of false alarms.

So as to guarantee the generalization performance, we also assume that hyper-ellipsoid centered at origin $\mathrm{E}(R,0)$ separates the two classes with margin $2d$, i.e. it satisfies the following constraints

$$\begin{aligned} R^2 - \|\mathbf{x}_i\|_M^2 &\ge d^2, \forall y_i{=}1 \\ \|\mathbf{x}_i\|_M^2 - R^2 &\ge d^2, \forall y_i{=}\text{-}1 \end{aligned} \tag{63}$$

where d is the shortest distance from the hyper-ellipsoid to the closest target and outlier class samples, $\|z\|_M := z^T \boldsymbol{\Sigma}^{-1} z$ for any vector z. Note that the distance is now under Mahalanobis distance metric and d act s just as the margin of the SVM.

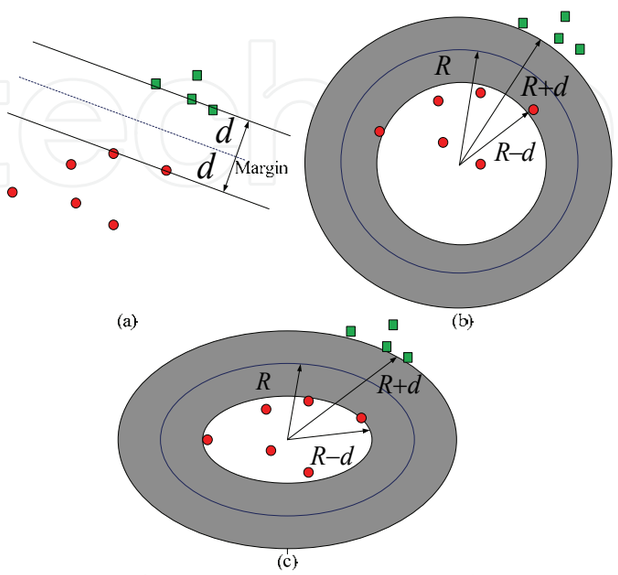


Fig. 1. Geometric illustrations of separation between two classes via different algorithms. (a) SVM. (b) Sphere shell separation (c) Ellipsoidal shell separation.

Obviously, there are many such hyper-ellipsoids which satisfy (63). An ideal criterion is to maximize the separation ratio $\frac{R+d}{R-d}$. But this objective is nonlinear and cannot be dealt with directly. Yet, it is easy to show that maximization of the separation ratio is equivalent to minimization of $\left(\frac{R}{d}\right)^2$. Using Taylor series formula, $\frac{R^2}{d^2}$ can be approximated as

$$\frac{R^2}{d^2} \approx \frac{R_0^2}{d_0^2} + \frac{1}{d_0^2}\left(R^2 - \frac{R_0^2}{d_0^2} d^2\right) \tag{64}$$

Now, the primal of Ellipsoidal shell separation in original space can be written as

$$\begin{aligned} &\min_{R^2, d^2} \quad R^2 - \gamma d^2 \\ &s.t. \quad y_i\left(R^2 - \|\mathbf{x}_i\|_M^2\right) \geq d^2 \end{aligned} \tag{65}$$

where γ is a constant, which controls the ratio of the radius to the separation margin.

4.2 The Linear Programming Classification Machine

Introducing the kernel trick, the primal form can be rewritten as follows:

$$\begin{aligned} &\min_{R^2,d^2} \quad R^2 - \gamma d^2 \\ &s.t. \quad y_i \left(R^2 - \sqrt{N} \left\| \mathbf{\Omega}^{-1} \mathbf{Q} \mathbf{k}_i \right\|^2 \right) \geq d^2 \end{aligned} \tag{66}$$

The robust kernelized primal version is

$$\begin{aligned} &\min_{R^2,d^2} R^2 - \gamma d^2 \\ &s.t. \ \ y_i \left(R^2 - \left\| (\frac{1}{N} \mathbf{\Omega} + r\mathbf{I})^{-\frac{1}{2}} \mathbf{\Omega}^{-\frac{1}{2}} \mathbf{Q} \mathbf{k}_i \right\|^2 \right) \geq d^2 \end{aligned} \tag{67}$$

So as to allow misclassified samples, we introduce slack variables $\xi_i \geq 0$. Thus (66) can be modified as

$$\begin{aligned} &\min_{R^2,d^2,\xi_i} R^2 - \gamma d^2 + C \sum_{i=1}^{N} \xi_i \\ &s.t. \begin{cases} y_i \left(R^2 - \left\| \mathbf{x}_i \right\|_M^2 \right) \geq d^2 - \xi_i \\ \xi_i \geq 0, i = 1, \cdots, N \end{cases} \end{aligned} \tag{68}$$

Accordingly, the kernelized primal version is

$$\begin{aligned} &\min_{R^2,d^2,\xi_i} R^2 - \gamma d^2 + C \sum_{i=1}^{N} \xi_i \\ &s.t. \begin{cases} y_i \left(R^2 - \sqrt{N} \left\| \mathbf{\Omega}^{-1} \mathbf{Q} \mathbf{k}_i \right\|^2 \right) \geq d^2 - \xi_i \\ \xi_i \geq 0, i = 1, \cdots, N \end{cases} \end{aligned} \tag{69}$$

And the robust kernel version is

$$\min_{R^2,d^2,\xi_i} R^2 - \gamma d^2 + C\sum_{i=1}^{N}\xi_i$$
$$s.t.\begin{cases} y_i\left(R^2 - \left\|(\frac{1}{N}\mathbf{\Omega} + r\mathbf{I})^{-\frac{1}{2}}\mathbf{\Omega}^{-\frac{1}{2}}\mathbf{Q}\mathbf{k}_i\right\|^2\right) \geq d^2 - \xi_i \\ \xi_i \geq 0, i = 1,\cdots,N \end{cases} \tag{70}$$

In order to obtain the obtain the dual form, the for the primal form (69) will be as follows:

$$L(R^2,d^2,\xi_i,\alpha_i,\beta_i) = R^2 - \gamma d^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(R^2 - \sqrt{N}\left\|\mathbf{\Omega}^{-1}\mathbf{Q}\mathbf{k}_i\right\|^2) - d^2 + \xi_i) - \sum_{i=1}^{N}\beta_i\xi_i \tag{71}$$

According to the KKT conditions, we will get following equalities:

$$\frac{\partial L}{\partial R^2} = 1 - \sum_{i=1}^{N}\alpha_i y_i = 0 \tag{72}$$

$$\frac{\partial L}{\partial d^2} = -\gamma + \sum_{i=1}^{N}\alpha_i = 0 \tag{73}$$

$$\frac{\partial L}{\partial \xi_i} = C - \beta_i - \alpha_i = 0 \tag{74}$$

Using (72)-(74), the Lagrange function for the primal form (69) is simplified into following form:

$$\min_{\alpha} -\sum_{i=1}^{N} \alpha_i y_i \sqrt{N} \left\| \mathbf{\Omega}^{-1} \mathbf{Q} \mathbf{k}_i \right\|^2$$
$$s.t. \begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 1 \\ \sum_{i=1}^{N} \alpha_i = \gamma \\ 0 \le \alpha_i \le C \\ i = 1, \cdots, N \end{cases} \tag{75}$$

We can see that (75) is in a linear programming form. Therefore, we can easily solve it using any mature and stable LP solvers. And it is expected to be easily extended to large scale datasets.
Accordingly, we can also conclude the dual form for (70) as follows:

$$\min_{\alpha} -\sum_{i=1}^{N} \alpha_i y_i \left\| (\frac{1}{N} \mathbf{\Omega} + r\mathbf{I})^{-\frac{1}{2}} \mathbf{\Omega}^{-\frac{1}{2}} \mathbf{Q} \mathbf{k}_i \right\|^2$$
$$s.t. \begin{cases} \sum_{i=1}^{N} \alpha_i y_i = 1 \\ \sum_{i=1}^{N} \alpha_i = \gamma \\ 0 \le \alpha_i \le C \\ i = 1, \cdots, N \end{cases} \tag{76}$$

5. Applications

5.1 Mahalanobis One Class SVMs

We investigate the initial performances of our proposed Mahalanobis Ellipsoidal Learning Machine (MELM) using three real-world datasets (ionosphere, heart and sonar) from the UCI machine learning repository. To see how well the MELM algorithm performs with respect to other learning algorithms, we compared the OCSVM, SVDD and MOCSVM algorithm using Gaussian kernels $k(x, y) = \exp(-\gamma \left\| x - y \right\|^2)$ .As for the MOCSVM, we only use one single RBF kernel for performance comparison.
We treat each class as the "normal" data in separate experiments. We randomly choose 80% of points as training data and the rest 20% as testing data. We determined the optimal

values of γ and C for RBF kernels by 5-fold cross validation. For the regularization constant, we set it as 0.01.
The datasets used and the results obtained by the four algorithms are summarized in Table 1. We can notice that the performance of our proposed Mahalanobis Ellipsoidal Learning Machine is competitive with or even better than the other approaches for the three datasets studied.

Dataset		SVDD	OCSVM	MOCSVM	MELM
ionosphere	+1	66.02	65.26	66.05	68.98
	-1	69.13	68.96	70.99	75.82
heart	+1	70.13	70.01	69.96	71.19
	-1	71.11	70.23	71.78	75.37
sonar	+1	92.98	92.10	93.29	96.49
	-1	89.73	89.38	90.49	94.32

Table 1. Performance results of different algorithms for single class problems. Correctness ratio (%) is reported.

From Table 1, we also see that, on all 3 datasets, the results obtained by the Mahalanobis distance based learning algorithm are slightly better than the corresponding results of the other two Euclidean distance based methods.

5.2 Mahalanobis Classification Machine

We tested the new algorithm and compared it to standard SVMs using several real-world datasets from the UCI machine learning repository. The results of MMEE and other three algorithms SVM, SSPC and MEME depend on the values of the kernel parameter λ and the regularization parameter C. In addition, the performance of MMEE, SSPC and MMEE also depend on the constant that balances the volume and the margin. For simplicity, we set C to infinity for all four algorithms. Thus, we only considered hard margin MMEE, SSPC, MEME and SVM algorithms.
For all the datasets, we used the 5-fold cross-validation method to estimate the generalization error of the classifiers. In the 5-fold cross-validation process, we ensured that each training set and each testing set were the same for all algorithms, and the same Gaussian kernel $k(x, y) = \exp(-\lambda \|x - y\|^2)$ was used. On each dataset, the value of the kernel parameter λ for SVM was optimized to provide the best error rate using 5-fold cross-validation. As for MMEE, We investigate the robust version MMEE. And the regularization constant r was set as 0.03. For SSPC, MEME and MMEE, the kernel parameter λ and the constant that balances the volume and the margin were optimized using grid based 5-fold CV method.
The datasets used and the results obtained by all four algorithms are summarized in Table 1. As we can see, SSPC and MEME achieve the same or slightly better results than SVMs on all 5 datasets. But our proposed MMEE method shows promising performances. The accuracy is commendably higher than the other three methods in the datasets studied in this paper.

Dataset	SVM	SSPC	MEME	MMEE
Breast Cancer	4.26	4.26	4.25	4.05
Ionosphere	6.00	5.71	5.70	4.84
Liver	36.19	35.36	35.42.	33.89
Pima	35.13	34.90	34.91	28.56
Sonar	11.22	10.73	11.03	9.93

Table 2. Performance results of different algorithms. Error rate (%) is reported.

6. Conclusions

In this paper, we extended the support vector data description one class support vector machines via utilizing the sample covariance matrix information and using the Mahalanobis distance metric instead of Euclidean distance metric. The proposed Mahalanobis Ellipsoidal Learning Machine can be easily addressed as a robust optimization problem by introducing an uncertainty model into the estimation of sample covariance matrix. We propose a LP representable Mahalanobis Data Description Machine for one class classification. We also address a robust optimization problem by introducing an uncertainty model into the estimation of sample covariance matrix. The results of applications to the three UCI real world datasets show promising performances.

We also proposed a LP based Minimum Mahalanobis Enclosing Ellipsoid (MMEE) pattern classification algorithm for generally two class dataset classification. The MMEE method can be solved in kernel form of LP. We also address a robust optimization problem by introducing an uncertainty model into the estimation of sample covariance matrix. Initial applications to several UCI real world datasets show promising performances. The initial results show that the proposed methods own both good description and discrimination character for supervised learning problems. Moreover, the data description with non-hyperplane bounding decision boundary owns better discrimination performance than hyperplane counterpart in the context of supervising learning.

7. References

Abe, Shigeo (2005). Training of support vector machines with Mahalanobis kernels. In *Proceeding of ICANN 2005 Conference*, LNCS 3697, W. Duch et al. Eds. Springer-Verlag, Berlin Heidelberg, pp. 571–576

Blake, C.; Keogh, E., Merz, C. J. (1998). UCI repository of machine learning databases, University of California, Irvine, CA. Available at http://www. ics.uci.edu/~ mlearn/ML-Repository.html

Chapelle, O. (2007). Training a Support Vector Machine in the Primal. *Neural Computation*, Vol.19, 1155-1178

Lanckriet, G.; El Ghaoui, L., Jodan M. (2003). Robust novelty detection with single-class MPM. In *Advances in Neural Information Processing Systems 15,* S.Becker, S. Thrun, and K. Obermayor, Eds. MIT Press, Cambridge

Li, Yinghong; Wei, Xunkai. (2004). *Engineering applications of support vector machines,* Weapon Industry Press, Beijing, China

Liu, Y.; Zheng, Y.-F. (2006). Maximum Enclosing and Maximum Excluding Machine for Pattern Description and Discrimination. In: Proceeding of ICPR 2006 Conference. Vol. 3, 129-132

Löfberg, J. (2004). YALMIP: A Toolbox for Modeling and Optimization in MATLAB. In: *Proceedings of the CACSD Conference*

Ruiz, A.; Lopez-de-Teruel, P. E. (2001). Nonlinear Kernel-based Statistical Pattern Analysis. *IEEE Transactions on Neural Networks,* Vol.2, No.1, 16-32

Scholkopf, B.; Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R. (2001). Estimating the Support of a High Dimensional Distribution. *Neural Computation*, Vol.13, No.7, 1443-1471

Tax, D.; Duin, R. (1999). Support Vector Domain Description. *Pattern Recognition Letters*, Vol.20, No.14, 1191-1199

Tsang, Ivor W.; Kwok, James T., and Li, Shutao. (2006). Learning the kernel in Mahalanobis one-class support vector machines. In *Proceeding of IJCNN 2006 Conference,* Vancouver, BC, Canada, 1169-1175

Wang, J.; Neskovic, P., Cooper, Leon N. (2005). Pattern Classification via Single Spheres. In: A. Hoffmann, H. Motoda, and T. Scheffer (eds.): *Proceeding of DS 2005 Conference*. LNAI, Vol. 3735, 241–252

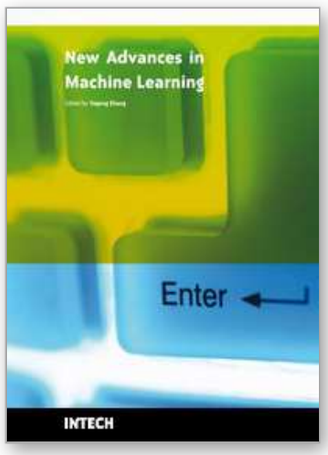
Wei, X.-K; Huang, G.-B., Li, Y.-H. (2007A). Mahalanobis Ellipsoidal Learning Machine for One Class Classification. In: *2007 International Conference on Machine Learning and Cybernetics*. Vol. 6, 3528-3533

Wei, X.-K.; Huang, G.-B., Li, Y.-H. (2007B). A New One Class Mahalanobis Hyperplane Learning Machine based on QP and SVD. In LSMS2007

Wei, X.-K.; Li, Y.-H., Feng, Y., Huang, G.-B. (2007C) Solving Mahalanobis Ellipsoidal Learning Machine via Second Order Cone Programming. In: De-Shuang Huang, et al. (eds.): Proceeding of ICIC 2007 Conference. CCIS, Vol.2, 1186-1194

Wei, X.-K.; Li, Y.-H., Li, Y.-F. (2007D). Enclosing Machine Learning: Concepts and Algorithms. *International Journal of Neural Computing and Applications*. Vol. 17, No. 3 , 237-243

Wei, X.-K.; Löfberg J., Feng, Y., Li, Y.-H, Li, Y.-F. (2007E). Enclosing Machine Learning for Class Description. In: D. Liu et al. (eds.): *Proceedings of ISNN2007 Conference*. LNCS, Vol. 4491, 428-437



New Advances in Machine Learning

Edited by Yagang Zhang

ISBN 978-953-307-034-6
Hard cover, 366 pages
Publisher InTech
Published online 01, February, 2010
Published in print edition February, 2010

The purpose of this book is to provide an up-to-date and systematical introduction to the principles and algorithms of machine learning. The definition of learning is broad enough to include most tasks that we commonly call “learning” tasks, as we use the word in daily life. It is also broad enough to encompass computers that improve from experience in quite straightforward ways. The book will be of interest to industrial engineers and scientists as well as academics who wish to pursue machine learning. The book is intended for both graduate and postgraduate students in fields such as computer science, cybernetics, system sciences, engineering, statistics, and social sciences, and as a reference for software professionals and practitioners. The wide scope of the book provides a good introduction to many approaches of machine learning, and it is also the source of useful bibliographical information.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen