On the Improvement of Support Vector Techniques for Clustering by Means of Whitening Transform

Stefanos Zafeiriou and Nikolaos Laskaris

***Abstract*—In this letter, we suggest a novel method for clustering, based on finding the smallest enclosing hyperellipse in arbitrary Hilbert spaces. In particular, we show that the one class support vector method that finds the minimum bounding hypersphere, under the whitening transform, becomes a method for finding the minimum bounding hyperellipse. Afterwards, we generalize the method in order to find the minimum bounding hyperellipse in arbitrary Hilbert spaces. We illustrate the power of the proposed methods in clustering applications.**

***Index Terms*—Clustering, kernel principal component analysis, support vector machines, whitening transform.**

I. Introduction

SUPPORT vector machines (SVMs) [1] have become synonym to ”good classifiers,” but their potential role in cluster analysis has been left relatively unexplored. The first method using SVM for clustering has been proposed in [2]. The so-called one class SVM method is used in order to model the underlying distribution of the data. The method proceeds by finding the hypersphere with the minimum radius in Hilbert spaces (usually using RBF kernels) that encloses all the training data [3]. The rest of SV clustering algorithms [2], [4], [5] work similarly and try to model the underlying distribution of the data in Hilbert spaces using hyperspheres. The hypersphere approximation in the Hilbert spaces is a valid assumption when all the features have equal variance in every dimension.

In this letter, we incorporate the statistics of the class distributions in one class SV formulations. To do so, we exploit the convenience of whitening-transform (WT) in order to build a new clustering method. The WT is one of the most commonly used normalizations in signal processing [6], [7] and is often related to hyperellipse modelling of distributions. With the proper algorithmic steps, the data, under the WT, are transformed in a format such that the variance in each dimension to be unity and their is no correlation between the dimensions (i.e., the covariance matrix of the data is the identity matrix). The goal of WT is either to decorrelate a data sequence prior to subsequent processing, or to control the spectral shape after processing. The interested reader may refer to [7] and to references therein for more details on the applicability of the WT.

We show that by incorporating WT in an SV method for finding the minimum bounding hypersphere of the data, a novel method emerges that can identify the minimum bounding hyperellipse. We propose a novel kernel method for estimating the underlying distribution of the data by identifying the minimum hyperellipse in Hilbert spaces that encloses the training data. We believe that the hyperellipse model is better suited for modelling the underlying cluster distribution, since the variance in each dimension is usually not the same in the input nor in the feature space. The hyperellipse modelling has received in the past much attention for clustering problems [8]–[13] where WT or similar approaches have been used. The proposed method converges to the hyperspherical SV methods when the variance is the same in every dimension in Hilbert space. Similar methods, to the proposed one, that show the benefits of hyperellipse modelling have been proposed [14]–[16]. Our method is considerably different from the optimization methods in [14]–[16] since we provide a robust solution in arbitrary dimensional Hilbert spaces and also a new clustering framework.

Experiments using synthetic and benchmark data accompany the theoretical exposition. To help the reader justify our suggestion, the results are presented in direct comparison with the results from standard SVM-techniques. The remainder of this letter is organized as follows. In Section II, the SV method for finding the minimum enclosing hyperellipse is presented. Experimental results are provided in Section III. Finally, conclusions are drawn in Section IV.

II. Hyperellipse Clustering

A. Whitening Transform

Let a training set with finite number of elements $\mathcal{X} = \{\mathbf{x}_i, i \in \{1, \ldots, N\}\}$, $\mathbf{x}_i \in \Re^M$. The total scatter matrix of the training data is

$$\mathbf{S}_t = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \mathbf{m})(\mathbf{x}_i - \mathbf{m})^T \tag{1}$$

where $\mathbf{m}$ is the mean vector of the set $\mathcal{X}$. Let also that the mean vector $\mathbf{m}$ is zero (i.e., we subtract the mean vector from all the training vectors). Assuming that the total scatter matrix $\mathbf{S}_t$ is invertible (i.e., the dimensionality of the samples $M \leq N$), the standard WT is a linear transform that takes the form

$$\mathbf{y}_i = \mathbf{S}_t^{-(1/2)}\mathbf{x}_i \tag{2}$$

and $\mathbf{S}_t^{-(1/2)} \in \Re^{M\times M}$, since $\mathbf{S}_t$ and $\mathbf{S}_t^{-1}$ are positive definite matrices. It can be easily verified that the total scatter matrix of vectors $\mathbf{y}_i$ is the $M \times M$ identity matrix $\mathbf{I}_M$.

B. Finding the Minimum Bounding Hyperellipsoid

In [2], an SV method has been proposed for finding the minimum bounding hypersphere. For the whitened vectors $\mathbf{y}_i$, given in (2), the optimization problem that finds the minimum bounding hypersphere is calculated by searching for the center $\mathbf{u}$ and the radius R derived from the following optimization problem:

$$\min R^2 \tag{3}$$

Manuscript received July 3, 2007; revised October 8, 2007. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Dominic Ho.

The authors are with the Aristotle University of Thessaloniki, Department of Informatics, 54124 Thessaloniki, Greece (e-mail: dralbert@aiia.csd.auth.gr; laskaris@aiia.csd.auth.gr).

Digital Object Identifier 10.1109/LSP.2007.914949

1070-9908/$25.00 © 2008 IEEE

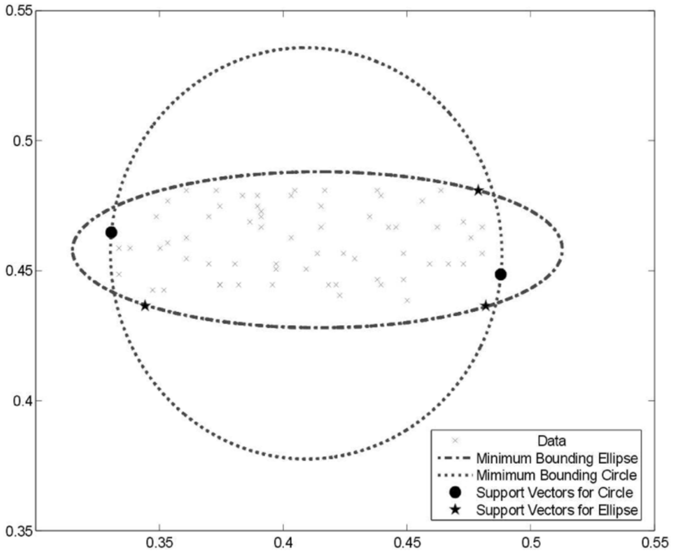

Fig. 1. Finding the minimum bounding circle and minimum bounding ellipse.

subject to the constraints

$$(\mathbf{y}_i - \mathbf{u})^T(\mathbf{y}_i - \mathbf{u}) \le R^2, \quad \forall\, \mathbf{y}_i \in \mathcal{X}. \qquad (4)$$

Now, substituting the whitened vectors $\mathbf{y}_i$ from (2)–(4), we take

$$\left(\mathbf{S}_t^{-(1/2)}\mathbf{x}_i - \mathbf{u}\right)^T \left(\mathbf{S}_t^{-(1/2)}\mathbf{x}_i - \mathbf{u}\right) \le R^2 \Leftrightarrow (\mathbf{x}_i - \mathbf{a})^T \mathbf{S}_t^{-1}(\mathbf{x}_i - \mathbf{a}) \le R^2, \quad \forall\, \mathbf{x}_i \in \mathcal{X} \qquad (5)$$

where $\mathbf{a} = \mathbf{S}_t^{(1/2)}\mathbf{u}$. This is equivalent to the problem of finding the minimum bounding hyperellipse.[1] Fig. 1 shows the minimum bounding sphere and the minimum bounding ellipse found using the proposed optimization problem (3) subject to the constraints (5). As can be seen, the WT leads to a better representation of the data than the representation based on the minimum bounding hypersphere.

Now we will generalize the optimization problem (3) subject to the (5) in arbitrary Hilbert spaces build using Mercer's kernels (especially using RBF kernels used in our experiments). To do so, let us define the nonlinear mapping $\phi : \Re^M \to \mathcal{H}$ that maps the training samples to the arbitrary dimensionally feature space. In this letter, we will restrict ourselves to the cases that the mapping ϕ satisfies the Mercer's condition [1]. The followings also hold for the linear case, and the simplistic choice $\phi(\mathbf{x}) = \mathbf{x}$. In all methods that will be described, the explicit closed form of the function ϕ is not needed, since for all the calculations, the dot products in $\mathcal{H}$ are employed using the so-called kernel trick as follows:

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j). \qquad (6)$$

The typical kernels that used in our experiments were RBF kernels

$$k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x})^T \phi(\mathbf{y}) = e^{-\gamma(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} \qquad (7)$$

where γ controls the spread of the Gaussian kernel.

Following the formulation of finding the smallest enclosing hypersphere in feature space [2], [17], the smallest enclosing hyperellipsoidal in feature space $\mathcal{H}$ is defined by optimizing (3) subject to the constraints

$$(\phi(\mathbf{x}_i) - \mathbf{a})^T \left(\mathbf{S}_t^{\Phi}\right)^{-1} (\phi(\mathbf{x}_i) - \mathbf{a}) \le R^2, \quad \forall\, \mathbf{x}_i \in \mathcal{X} \qquad (8)$$

where $\mathbf{S}_t^{\Phi}$ is the covariance matrix defined in arbitrary Hilbert spaces as

$$\mathbf{S}_t^{\Phi} = \frac{1}{N} \sum_{\mathbf{x} \in \mathcal{X}} (\phi(\mathbf{x}) - \mathbf{m}^{\Phi})(\phi(\mathbf{x}) - \mathbf{m}^{\Phi})^T. \qquad (9)$$

There is a major difficulty in finding the minimum bounding hyperellipsoidal in Hilbert spaces. The difficulty is that we could not satisfy the invertibility of $\mathbf{S}_t^{\Phi}$ (the rank of $\mathbf{S}_t^{\Phi}$ is smaller than N) in arbitrary dimensional Hilbert spaces, and we need a closed form solution for the matrix $\left(\mathbf{S}_t^{\Phi}\right)^{-1}$. That is, we need to calculate an approximation matrix $\mathbf{W}^{\Phi}$ of the matrix $\mathbf{S}_t^{\Phi}$. The properties of the matrix $\mathbf{W}_t^{\Phi}$ are listed below.

- It keeps the principal structure of the covariance matrix $\mathbf{S}_t^{\Phi}$, so as the dominant eigenvalues and eigenvectors of $\mathbf{S}_t^{\Phi}$ and $\mathbf{W}_t^{\Phi}$ remain the same.
- It is compact and regularized. The compactness is inspired by the fact that the smallest eigenvalues of the covariance matrix are very close to zero. The regularity is always desirable in the approximation theory.
- It is easy to invert analytically. This is especially important in our case since we want a closed form expression of $\mathbf{S}_t^{\Phi^{-1}}$.

Such a solution has been given in [18] as

$$\mathbf{W}_t^{\Phi} = \mathbf{\Phi}\mathbf{J}\mathbf{Q}\mathbf{Q}^T\mathbf{J}^T\mathbf{\Phi}^T + \rho\mathbf{I}^{\Phi} \qquad (10)$$

where $\mathbf{J} = (1/\sqrt{N})\left(\mathbf{I}_N - (1/N)\mathbf{1}\mathbf{1}^T\right)$, $\mathbf{I}^{\Phi}$ is the identity matrix in the feature space, $\mathbf{1}$ is a N-dimensional vector of ones, $\mathbf{\Phi} = [\phi(\mathbf{x}_1), \ldots, \phi(\mathbf{x}_N)]$, and ρ is a constant used for regularization (i.e., the typical value is 10^{-3}). In order to define $\mathbf{Q}$, we have first to define the centralized Gram matrix $\bar{\mathbf{K}}$ as

$$\bar{\mathbf{K}} = \mathbf{J}^T\mathbf{\Phi}^T\mathbf{\Phi}\mathbf{J} \qquad (11)$$

the Gram matrix $\bar{\mathbf{K}}$ is $N \times N$. Thus, it is computational feasible to perform eigenanalysis to $\bar{\mathbf{K}}$ and let $\{\lambda_n, \mathbf{v}_n\}_{n=1}^{r}$ be the corresponding set of eigenvalues and eigenvectors, respectively. The previous procedure for finding the r top eigenvectors that correspond to non-zero eigenvalues of the matrix $\bar{\mathbf{K}}$ is the KPCA transform [19]. The matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \mathbf{V}\left(\mathbf{I}_r - \rho\mathbf{L}_r^{-1}\right)^{(1/2)} \qquad (12)$$

where $\mathbf{L}_r = \mathrm{diag}[\lambda_1, \ldots, \lambda_r]$ is a diagonal matrix with elements the non-null eigenvalues of $\bar{\mathbf{K}}$, and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_r]$ is the matrix with columns the corresponding eigenvectors.

Now the closed form for $\mathbf{W}_t^{\Phi^{-1}}$ is given by the Woodbury formula [18] as

$$\mathbf{W}_t^{\Phi^{-1}} = \rho^{-1}(\mathbf{I}^{\Phi} - \mathbf{\Phi}\mathbf{B}\mathbf{\Phi}) \qquad (13)$$

where

$$\mathbf{B} = \mathbf{J}\mathbf{Q}\mathbf{M}^{-1}\mathbf{Q}^T\mathbf{J}^T \qquad (14)$$

[1] To be strict, in this letter, we define the hyperellipses using the total scatter matrix, which is an approximation of finding the true minimum volume hyperellipse. For finding the "true" hyperellipse, one should initially solve an optimization problem for finding the actual covariance problem.

and the matrix $\mathbf{M}$ can be thought of as a "reciprocal" matrix of $\mathbf{W}_t^{\Phi}$

$$\mathbf{M} = \rho \mathbf{I}_r + \mathbf{Q}^T \mathbf{J}^T \mathbf{\Phi}^T \mathbf{\Phi} \mathbf{J} \mathbf{Q}. \tag{15}$$

Having found an approximation of the matrix $\mathbf{S}^{\Phi}$, the matrix $\mathbf{W}^{\Phi}$, that meets all the necessary requirements mentioned above, we can proceed in solving the quadratic optimization problem that will give us the center $\mathbf{a}$ and radius R. In case that we take into consideration the presence of outliers, the optimization problem (3) subject to the constraints (8) is reformulated in its *soft* form as

$$\min R^2 + C \sum_{i=1}^{N} \xi_i \tag{16}$$

subject to

$$(\phi(\mathbf{x}_i) - \mathbf{a})^T \mathbf{W}_t^{\Phi^{-1}} (\phi(\mathbf{x}_i) - \mathbf{a}) \leq R^2 + \xi_i, \quad \forall\, \mathbf{x}_i \in \mathcal{X} \tag{17}$$

with slack variables $\xi_i \geq 0, \forall\, i = 1, \ldots, N$. We solve the problem of finding the smallest bounding hyperellipse by founding the saddle point of the Lagrangian

$$\begin{aligned} L(R, \mathbf{a}, \boldsymbol{\xi}) = R^2 - \sum_{i=1}^{N} \Bigg(& R^2 + \xi_i - (\phi(\mathbf{x}_i) - \mathbf{a})^T \mathbf{W}_t^{\Phi^{-1}} \\ & \times (\phi(\mathbf{x}_i) - \mathbf{a}) \Bigg) \alpha_i - \sum_{i=1}^{N} \xi_i \beta_i + C \sum_{i=1}^{N} \xi_i \end{aligned} \tag{18}$$

where $\alpha_i \geq 0$ and $\beta_i \geq 0$ for $i = 1, \ldots, N$ are the Lagrange multipliers associated to the constraints (18), and $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_N]$ is the vector of slack variables. The conditions that should be met for the saddle point are

$$\begin{aligned} \nabla_{\mathbf{a}} L|_{\mathbf{a}=\mathbf{a}_o} &= \mathbf{0} \Leftrightarrow \mathbf{a}_o = \frac{1}{2} \sum_{i=1}^{N} a_{i,o} \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}_i) \\ \frac{\partial L}{\partial R}|_{R=R_o} &= 0 \Leftrightarrow \mathbf{a}_o^T \mathbf{1} = 1 \\ \frac{\partial L_3}{\partial \xi_i}|_{\xi_i=\xi_{i,o}} &= 0 \Leftrightarrow b_{i,o} = C - a_{i,o} \end{aligned} \tag{19}$$

substituting the above conditions to (18), the Wolfe dual form of the constraint optimization problem is the maximization of

$$\begin{aligned} W(\boldsymbol{\alpha}) = & \sum_{i=1}^{N} a_{i,o} \phi(\mathbf{x}_i)^T \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}_i) \\ & - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_{i,o} a_{j,o} \phi(\mathbf{x}_i)^T \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}_j) \end{aligned} \tag{20}$$

subject to the constraints

$$0 \leq a_{i,o} \leq C \text{ and } \sum_{i=1}^{N} a_{i,o} = 1, \quad \forall\, i = 1, \ldots, N. \tag{21}$$

The way the term $\phi(\mathbf{x}_i)^T \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}_j)$ is calculated using the so-called kernel trick is shown in the Appendix. The Karush–Kuhn–Tucker (KKT) conditions of the optimization problem (16) subject to the constraints (17) yield

$$\begin{aligned} \left(R^2 + \xi_i - (\phi(\mathbf{x}_i) - \mathbf{a})^T \mathbf{W}_t^{\Phi^{-1}} (\phi(\mathbf{x}_i) - \mathbf{a})\right) \alpha_i = 0 \\ \xi_i \beta_i = 0, \quad \forall\, i = 1, \ldots, N. \end{aligned} \tag{22}$$

From the KKT conditions, it is clear that only the vectors $\phi(\mathbf{x}_i)$ with $a_{i,o} \neq 0$ are needed for defining the center of the hyperellipsoidal in $\mathcal{H}$ (support vectors). We can further compute the distance between input pattern and the center using the distance

$$\begin{aligned} D^2(\mathbf{x}) &= (\phi(\mathbf{x}) - \mathbf{a}) \mathbf{W}_t^{\Phi^{-1}} (\phi(\mathbf{x}) - \mathbf{a}) \\ &= \phi(\mathbf{x})^T \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}) - 2 \sum_i \beta_i \phi(\mathbf{x}_i)^T \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}) \\ &\quad + \sum_i \sum_j \phi(\mathbf{x}_i)^T \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}_j) \end{aligned} \tag{23}$$

which is actually the Mahalanobis distance in Hilbert space $\mathcal{H}$. The best R is the one

$$R = \{\sqrt{D(\mathbf{x}_i)^2} : \mathbf{x}_i \text{ is an SV}\}. \tag{24}$$

The point $\phi(\mathbf{x}_i)$ is outside the minimum bounding hyperellipse if the corresponding slack variable $\xi_i > 0$. From the KKT conditions, we know that $\beta_i = C$. These vectors are the outer vectors. These vectors do not exist when $C \geq 1$. Furthermore, the point is located in the hyperellipsoidal surface when $0 < \beta_i < C$. The points $\phi(\mathbf{x}_i)$ lie inside the hyperellipse when $\beta_i = 0$. The method for finding the minimum bounding hyperellipse can be converted to a clustering method using one of the methods [2], [4], [5]. We applied the proposed method in the framework proposed [5]. This clustering framework is described in more detail in [20]. This method can be interpreted as an extension of k-means clustering in Hilbert spaces (using Mercer's kernels). That is, the data are clustered in k hyperspheres (in our method, these are hyperellipses) in feature space. The center and the shape of each cluster is defined by each of the hyperspheres. In our case, the shape and the center of the cluster is defined by each hyperellipse. All the SV-based clustering approaches [2], [4], [5] use hyperspheres in order to model the underlying distribution of the data (or the distribution of each cluster). The hypersphere distribution approximation in the feature and in the input space is a valid assumption when all the features are of equal variance. We believe that the hyperellipse approximation for each cluster is more valid due to the fact that it allows the data to have different variances in every dimension.

III. Clustering Experiments in Various Datasets

Our algorithm of hyperellipsoidal clustering has been tried on three data sets, that is, the IRIS Database [21], the Spam data set [22], and the Wisconsin's breast cancer database [23]. We have compared the proposed clustering scheme in terms of performance with an implementation of the hyperspherical SVM-based clustering scheme in [5]. The proposed scheme has all the advantages of SVM-based implementation proposed in [5]. Thus, we have not conducted experiments with other clustering schemes like K-Means, Neural Gas, SOM, and Ng-Jordan algorithm [24] since, in [5], it has been shown that these algorithms are constantly outperformed by the SVM-based clustering algorithm in [5]. Nevertheless, we have performed experiments using an implementation of the hyperellipse clustering in [11].

We tested our method on the IRIS data using one center for each of the three classes. The results obtained using the hyperspherical SVM clustering [5] and the proposed hyperellipsoidal SVM clustering for the IRIS database (150 samples) are summarized in column 2 of Table I (correctly classified samples). The Spam database collects 1534 samples from two different classes, spam and not-spam. Each sample is represented by a

TABLE I
HYPERSPHERICAL SV CLUSTERING VERSUS HYPERELLIPSOIDAL SV CLUSTERING IN IRIS, SPAM, AND WISCONSIN DATABASE

Algorithm	IRIS Data	Spam Database	Wisconsin Cancer database
Hyperspherical SV Clustering [5]	142 ± 1 (94.7%)	1247 ± 3 (81.3%)	662.5 ± 0.5 (97%)
Hypellipsoidal Clustering [11]	123 ± 1 (82%)	1220 ± 3 (79.5%)	664 ± 3 (97.2%)
Hyperellipsoidal SV Clustering	**145.5** ± 1.5 (97%)	**1310** ± 5 (85.4%)	**668.5** ± 1 (98%)

57-dimensional feature vector. The third column of Table I displays the average performances on 20 runs obtained on Spam database for different initializations and parameters for both the hyperspherical and hyperellipsoidal clustering. Last, we have experimented with the Wisconsin's breast cancer database [23]. This database contains 683 nine-dimensional samples. The clustering results obtained in this database are summarized in column 4 of Table I. We have compared the proposed clustering model with the model proposed hyperspherical model in [5] and with an implementation of the hyperellipse clustering in [11]. In Table I, we have included the clustering results from [5], since in our experiments, the hyperspherical algorithm showed slightly different performance. It can be seen that our method obtains consistently better results than the hyperspherical SVM clustering and the an implementation of the clustering in [11].

In the experiments presented, typical eigenanalysis routines have been used. Since, the eigenanalysis of the total scatter matrix is one of the steps of the proposed approach that may require extra care in order to avoid numerical instability and the effect of outliers. Some methods that can be used for such purposes can be found in [25] and more recently in [26]. From our experience, the performance can be increased when, apart from the eigenvectors that correspond to null eigenvalues, some of the eigenvectors that correspond to small eigenvalues are discarded, as well.

IV. CONCLUSIONS

In this letter, we have shown how SV methods for clustering are transformed under the whitening normalization. That is, we have developed a method for finding the minimum bounding hyperellipse. Experimental results proved that the whitening normalization may serve in improving the performance of the SV method. A possible disadvantage of the proposed method is that it requires an initial eigenanalysis of the total scatter matrix. Further research on the topic for a practitioner's perspective is the incorporation of robust statistics [27], [28] for the calculation of the total scatter matrix in order to cope with the presence of possible outliers in the class distributions.

APPENDIX
CALCULATING $\phi(\mathbf{x}_i)^t \mathbf{W}_t^{\phi^{-1}} \phi(\mathbf{x}_j)$

The term $\phi(\mathbf{x}_i)^T \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}_j)$ can be represented in terms of vector dot products (kernel-trick) in the Hilbert space $\mathcal{H}$ as follows:

$$\begin{aligned}\phi(\mathbf{x}_i)^T \mathbf{W}_t^{\Phi^{-1}} \phi(\mathbf{x}_j) &= \phi(\mathbf{x}_i)^T \rho^{-1}(\mathbf{I}^\Phi - \mathbf{\Phi}\mathbf{B}\mathbf{\Phi}^T)\phi(\mathbf{x}_j)\\ &= \rho^{-1}k(\mathbf{x}_i,\mathbf{x}_j)\\ &\quad - \rho^{-1}\phi(\mathbf{x}_i)^T \mathbf{\Phi}\mathbf{B}\mathbf{\Phi}^T \phi(\mathbf{x}_j)\\ &= \rho^{-1}k(\mathbf{x}_i,\mathbf{x}_j)\\ &\quad - \rho^{-1}[k(\mathbf{x}_1,\mathbf{x}_i),\ldots,k(\mathbf{x}_N,\mathbf{x}_i)]\\ &\quad \times \mathbf{B}[k(\mathbf{x}_1,\mathbf{x}_j),\ldots,k(\mathbf{x}_N,\mathbf{x}_j)]^T. \end{aligned} \tag{25}$$

REFERENCES

[1] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[2] A. Ben-Hur, D. Horn, H. T. Siegelmann, and V. Vapnik, "Support vector clustering," *J. Mach. Learn. Res.*, vol. 2, pp. 125–137, 2001.

[3] C. J. C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mine. Knowl. Discov.*, vol. 2, pp. 121–167, 1998.

[4] J.-H. Chiang and P.-Y. Hao, "A new kernel-based fussy clustering approach: Support vector clustering with cell growing," *IEEE Trans. Fuzzy Syst.*, vol. 11, no. 3, pp. 518–527, Aug. 2003.

[5] F. Camastra and A. Verri, "A novel kernel method for clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 5, pp. 801–805, May 2005.

[6] K. Fukunaga, *Statistical Pattern Recognition*. San Diego, CA: Academic, 1990.

[7] Y. C. Eldar and A. V. Oppenheim, "MMSE whitening and subspace whitening," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1846–1851, Jul. 2003.

[8] E.E. Gustafson and W.C. Kessel, "Fuzzy clustering with fuzzy covariance matrix," in *Proc. IEEE CDC*, San Diego, CA, 1979, pp. 765–784.

[9] S. N. Kavuri and V. Venkatasubramanian, "Using fuzzy clustering with ellipsoidal units in neural networks for robust fault classification," *Comput. Chem. Eng.*, vol. 17, no. 8, pp. 765–784, 1993.

[10] R. Krishnapuram, H. Frigui, and O. Nasraoui, "Fuzzy and possibilistic shell clustering algorithms and their application to boundary detection and surface approximation—parts i and ii," *IEEE Trans. Fuzzy Syst.*, vol. 3, no. 1, pp. 29–60, Feb. 1995.

[11] J. Mao and A. Jain, "A self-organizing network for hyperellipsoidal clustering (hec)," *IEEE Trans. Neural Netw.*, vol. 7, no. 1, pp. 16–29, Jan. 1996.

[12] W. Song, X. Shaowei, J. Mao, A. Jain, D. V. Prokhorov, and D. C. Wansch, II, "Comments on "A self-organizing network for hyperellipsoidal clustering (hec)"," *IEEE Trans. Neural Netw.*, vol. 8, no. 6, pp. 1561–1563, Nov. 1997.

[13] C. F. Juang and C. T. Lin, "An on-line self-constructing neural fuzzy inference network and its applications," *IEEE Trans. Fuzzy Syst.*, vol. 6, no. 1, pp. 12–32, Feb. 1998.

[14] P. Sun and R. M. Freund, "Computation of minimum-volume covering ellipsoids," *Oper. Res.*, vol. 52, no. 5, pp. 690–706, 2004.

[15] D. M. J. Tax and P. Juszczak, "Kernel whitening for one-class classification," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 17, no. 3, pp. 333–347, 2003.

[16] F. Glineur, "Pattern separation via ellipsoids and conic programming," Master's thesis, Faculte Polytechnique de Mons, Mons, Belgium, 1998.

[17] B. Scholkopf, A. Williamson, A. J. Smola, J. Shawe-Taylor, and J. Platt, "Support method for novelty detection," *Adv. Neural Inf. Process. Syst.*, vol. 12, pp. 582–588, 2000.

[18] S. K. Zhou and R. Chellappa, "From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel Hilbert space," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 917–929, Jun. 2006.

[19] A. Scholkopf, B. Smola, and K. R. Muller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, pp. 1299–1319, 1998.

[20] F. Camastra, "Kernel methods for unsupervised learning," Ph.D. dissertation, Univ. Genova, Genova, Italy, 2004.

[21] IRIS database. [Online]. Available: ftp.ics.uci.edu/pub/machine-learning-databases/iris,

[22] SPAM database. [Online]. Available: ftp.ics.uci.edu/pub/machine-learning-databases/spam.

[23] W. H. Wolberg and O. L. Mangasarian, "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Proc. Nat. Acad. Sci.*, vol. 87, pp. 9193–9196, 1990.

[24] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Proc. Advances in Neural Information Processing Systems (NIPS 2001)*, 2001, vol. 14, pp. 849–856.

[25] L. Juwei, K. N. Plataniotis, and A. N. Venetsanopoulos, "Regularization studies of linear discriminant analysis in small sample size scenarios with application to face recognition," *Pattern Recognit. Lett.*, vol. 26, no. 2, pp. 181–191, 2005.

[26] J. Xudong, M. Bappaditya, and K. Alex, "Eigenfeature regularization and extraction in face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, accepted for publication.

[27] G. Seber, *Multivariate Observations*. New York: Wiley, 1986.

[28] P. J. Huber, *Robust Statistics*. New York: Wiley, 1981.