Mercer Kernel-Based Clustering in Feature Space

Mark Girolami

***Abstract*—This letter presents a method for both the unsupervised partitioning of a sample of data and the estimation of the possible number of inherent clusters which generate the data. This work exploits the notion that performing a nonlinear data transformation into some high dimensional feature space increases the probability of the linear separability of the patterns within the transformed space and therefore simplifies the associated data structure. It is shown that the eigenvectors of a kernel matrix which defines the implicit mapping provides a means to estimate the number of clusters inherent within the data and a computationally simple iterative procedure is presented for the subsequent feature space partitioning of the data.**

***Index Terms*—Data clustering, data partitioning, unsupervised learning.**

I. Introduction

The unsupervised partitioning of a sample of data observations into self-similar regions forms a significant area of research effort. As it has been noted that many data sets have ellipsoidal clustered structure "sum-of-squares" based methods of partitioning have proved to be effective [4]. Clustering using Gaussian mixture models is also extensively employed for exploratory data analysis. However, in certain cases the number of Gaussian mixtures required to reasonably model the data density far exceeds the natural number of clusters in the data. This is of course the case when the clusters themselves are non-Gaussian [7].

For the purposes of classification the problem of nonlinear separability of classes can be circumvented by mapping the observed data to a higher dimensional space in a nonlinear manner so that each cluster for each class unfolds into a simple form. This is the basis for nonlinear classification techniques such as radial basis function networks, support vector (SV) machines [11], and certain forms of nonlinear discriminant analysis [9]. If the nonlinear mapping is smooth and continuous then the topographic ordering of the data in observation space will be preserved in feature space, so that points clustered together in data space will necessarily be clustered in feature space. It is therefore of interest to consider the further notion of unsupervised data clustering in a feature space which preserves the inherent data groupings and in addition simplifies the associated structure of the data.

Section II reconsiders sum-of-squares clustering in data space while Section III presents the clustering of data in a nonlinear feature space. Section IV of this paper considers how the block diagonal structure of a kernel matrix can be exploited in estimating the number of inherent clusters within a data sample. Section V provides some demonstrative simulations and Section VI provides conclusions and discussion.

II. A Data-Space Clustering Criterion

The sum-of-squares cost for a sample of data forms the basis for a number of clustering methods [4], [1]. Given a finite set of observations of datum vector $\mathbf{x}_n$; $n = 1, \ldots, N$ where $\mathbf{x}_n \in \mathbf{R}^D$ and given K cluster centers the within-group scatter matrix is defined as

$$\mathbf{S}_W = \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{kn}(\mathbf{x}_n - \mathbf{m}_k)(\mathbf{x}_n - \mathbf{m}_k)^T \tag{1}$$

where the center for each group is defined as $\mathbf{m}_k = N_k^{-1}\sum_{n=1}^{N} z_{kn}\mathbf{x}_n$ and $N_k = \sum_{n=1}^{N} z_{kn}$. The variable z_{kn} indicates the membership of datum $\mathbf{x}_n$ to cluster k, i.e., $z_{kn} = 1$ if $\mathbf{x}_n \in C_k$ and 0 otherwise. One of the sum-of-squares criteria employed for central clustering is the trace of the within-group scatter matrix $\mathrm{Tr}(\mathbf{S}_W)$. This measure implicitly imposes the assumption of hyper-spherical clusters which is inherent in methods such as the K-means algorithm [4]. The $K \times N$ indicator matrix $\mathbf{Z}$ is such that each element is either of two values 1 or 0, such that $z_{ki} \in \{0, 1\}$ $\forall k, i$ and $\sum_{k=1}^{K} z_{ki} = 1$ $\forall i$. The optimal partitioning of the data sample is achieved by the following optimization:

$$\mathbf{Z} = \arg\min_{\mathbf{Z}} \mathrm{Tr}(\mathbf{S}_W). \tag{2}$$

Methods such as the K-means algorithm and its many variants are used in the optimization of the above data space sum-of-squares clustering criterion [4]. If the separation boundaries between clusters is nonlinear then sum-of-squares methods such as K-means will fail. Semiparametric mixture-decomposition methods such as the recently developed maximum-certainty partitioning [7] have been proposed to deal with the problem of non-Gaussian clustered data. An alternative approach to solving this problem is to adopt the strategy of nonlinearly transforming the data into a high-dimensional feature space and then performing the clustering within this feature space. However as the feature space may be of high and possibly infinite dimension then directly working with the transformed variables is an unrealistic option. However, as has been exploited in the kernel principal component analysis (KPCA) method of feature extraction it is unnecessary to work directly with the transformed variables [10]. It is the inner-products between points which are used and these can be computed using a kernel function in the original data space. This observation provides for a tractable means of working in the possibly infinite feature spaces [11], [10]. We now develop the feature space sum-of-squares clustering method in the following section.

III. Feature Space Clustering

The implicit assumption of hyper-spherical or hyper-ellipsoidal clusters is often restrictive and, similar to classification problems, a nonlinear mapping into some higher dimensional space which will provide linear separation of classes is desirable [11]. Consider then a smooth, continuous nonlinear mapping from data space to feature space F such that

$$\Phi: \mathbf{R}^D \longrightarrow F \qquad \mathbf{x} \mapsto \mathbf{X}.$$

Denoting the within-group scatter matrix in feature space F as $\mathbf{S}_W^{\Phi}$ then the trace of the feature space scatter matrix is given by

$$\begin{aligned}\mathrm{Tr}\left(\mathbf{S}_W^{\Phi}\right) &= \mathrm{Tr}\left\{\frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{kn}\left(\Phi(\mathbf{x}_n) - \mathbf{m}_k^{\Phi}\right)\left(\Phi(\mathbf{x}_n) - \mathbf{m}_k^{\Phi}\right)^T\right\} \\ &= \frac{1}{N}\sum_{k=1}^{K}\sum_{n=1}^{N} z_{kn}\left(\Phi(\mathbf{x}_n) - \mathbf{m}_k^{\Phi}\right)^T\left(\Phi(\mathbf{x}_n) - \mathbf{m}_k^{\Phi}\right). \end{aligned} \tag{3}$$

The cluster center in feature space is now denoted by the following expression $\mathbf{m}_k^{\Phi} = N_k^{-1}\sum_{n=1}^{N} z_{kn}\Phi(\mathbf{x}_n)$. It is interesting, and fortu-

Manuscript received January 8, 2001; revised February 21, 2001. This work was supported by the Council for Museums Archives and Libraries, Grant RE/092 "Improved Online Information Access" and in part by the Finnish National Technology Agency TEKES.

The author is with the Laboratory of Computing and Information Science, Helsinki University of Technology, FIN-02015 HUT, Hensinki, Finland, on Secondment from Applied Computational Intelligence Research Unit, Department of Computing and Information Systems, University of Paisley, U.K. (e-mail: mark.girolami@paisley.ac.uk).

Publisher Item Identifier S 1045-9227(02)05003-8.

itous, to note that $\mathrm{Tr}(\mathbf{S}_W^{\Phi})$ takes the form of a series of dot products in feature space. As noted in the previous section these feature space dot products can easily be computed using Mercer kernel [11], [10] representations in data space $k(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$. In other words by employing a specific kernel function the dot product which it returns implicitly defines the nonlinear mapping Φ to the feature space [11], [10]. The feature space sum-of-squares criterion can now be written solely in terms of elements of the symmetric $N \times N$ kernel matrix $\mathbf{K} = \{K_{ij}\}_{i=1,\ldots,N;\, j=1,\ldots,N}$ where $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \equiv \Phi(\mathbf{x}_i).\Phi(\mathbf{x}_j)$ and $K_{ij} = K_{j\,i}$

$$\mathrm{Tr}\left(\mathbf{S}_W^{\Phi}\right) = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} z_{kn} y_{kn} \tag{4}$$

where

$$y_{kn} = K_{nn} - \frac{2}{N_k} \sum_{j=1}^{N} z_{kj} K_{nj} + \frac{1}{N_k^2} \sum_{i=1}^{N} \sum_{l=1}^{N} z_{ki} z_{kl} K_{il}. \tag{5}$$

By defining the following terms $\gamma_k = N_k/N$ and, $\mathcal{R}(\mathbf{x}|C_k) = N_k^{-2} \sum_{i=1}^{N} \sum_{j=1}^{N} z_{ki} z_{kj} K_{ij}$ where the notation $\mathcal{R}(\mathbf{x}|C_k)$ denotes the quadratic sum of the elements which have been allocated to the kth cluster, then some straightforward manipulation of (4) yields

$$\mathrm{Tr}\left(\mathbf{S}_W^{\Phi}\right) = \frac{1}{N} \sum_{k=1}^{K} \sum_{n=1}^{N} z_{kn} K_{nn} - \sum_{k=1}^{K} \gamma_k \mathcal{R}(\mathbf{x}|C_k). \tag{6}$$

For kernels which depend on the difference $(\mathbf{x}_i - \mathbf{x}_j)$ then the first term in (6) will be a constant, indeed for the widely used RBF kernel $k(\mathbf{x}_i, \mathbf{x}_j) = \exp\{-(1/c)\|\mathbf{x}_i - \mathbf{x}_j\|^2\}$ the first term equals unity thus (6) reduces to

$$\mathrm{Tr}\left(\mathbf{S}_W^{\Phi}\right) = 1 - \sum_{k=1}^{K} \gamma_k \mathcal{R}(\mathbf{x}|C_k). \tag{7}$$

The implicit assumption of hyper-spherical clusters in the sum-of-squares criterion is now based on the feature space representation of the data which is defined by the specific kernel chosen. The RBF kernel implicitly defines an infinite dimensional feature space, this particular kernel has been extensively adopted in many studies of both classification [11] and unsupervised learning [10].

If we now consider the RBF kernel specifically it is straightforward to see that as $\sum_k z_k = 1$ then $0 < \mathcal{R}(\mathbf{x}|C_k) \leq 1$. In addition as $\sum_k \gamma_k = 1$ then $\sum_{k=1}^{K} \gamma_k \mathcal{R}(\mathbf{x}|C_k) \leq 1$ in which case the minimization of $\mathrm{Tr}(\mathbf{S}_W^{\Phi})$ requires the maximization of $\sum_{k=1}^{K} \gamma_k \mathcal{R}(\mathbf{x}|C_k)$. It is worthy of note that for an RBF kernel the following approximation, which was originally utilized in [2], holds due to the convolution theorem for Gaussians

$$\int_{\mathbf{x}} p(\mathbf{x})^2 \, d\mathbf{x} \approx \frac{1}{N^2} \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij}. \tag{8}$$

This being the case then

$$\mathcal{R}(\mathbf{x}|C_k) = \frac{1}{N_k^2} \sum_{i=1}^{N} \sum_{j=1}^{N} z_{ki} z_{kj} K_{ij} \approx \int_{\mathbf{x} \in C_k} p(\mathbf{x}|C_k)^2 \, d\mathbf{x}.$$

So the term defined as $\mathcal{R}(\mathbf{x}|C_k)$ can be considered as a nonparametric approximation to the integral $\int_{\mathbf{x} \in C_k} p(\mathbf{x}|C_k)^2 d\mathbf{x}$ defined over the kth cluster. As already stated this was originally proposed in [2] as a measure of distribution compactness based on a nonparametric estimate of the probability density of the data. So in this particular case $\mathcal{R}(\mathbf{x}|C_k)$ provides a measure of the compactness of the kth cluster as defined above. This is in contrast to the Euclidean compactness measure defined by the sum-of-squares error computed in the original data space given by equation (2). The desired sum-of-squares clustering in a feature space defined by an RBF kernel is therefore represented by the following nonlinear optimization problem:

$$\mathbf{Z} = \arg\min_{\mathbf{Z}} \mathrm{Tr}\left(\mathbf{S}_W^{\Phi}\right) = \arg\max_{\mathbf{Z}} \sum_{k=1}^{K} \gamma_k \mathcal{R}(\mathbf{x}|C_k). \tag{9}$$

What becomes clear is that feature space clustering achieved by employing a kernel representation of the data removes the implicit assumption of hyper-spherical or ellipsoidal clusters in data space. For the particular case of the popular and widely used RBF kernel then the optimization of the feature space criterion given by equation (9) is required. In considering the optimization of the clustering criterion (9) it is proposed that the following lemma, originally detailed in [3], is utilized.

Lemma: If the restriction $z_{ki} \in \{0, 1\}\ \forall k,\ i$ is relaxed to $0 \leq z_{ki} \leq 1\ \forall k, i$, i.e., $z_{ki} \in [0, 1]$ with the summation constraint holding then the minimum of a sum of squares clustering criterion ($\mathrm{Tr}(\mathbf{S}_W)$ or $\mathrm{Tr}(\mathbf{S}_W^{\Phi})$) is achieved with a matrix $\mathbf{Z}$ which has elements zero or one only.

The complete proof is given in [3]. This lemma also has a probabilistic interpretation in that the maximum certainty partitioning of data will only occur when the partition posteriors are zero or one [7]. This important lemma provides for the use of stochastic methods in optimizing clustering criteria based on a binary indicator matrix.

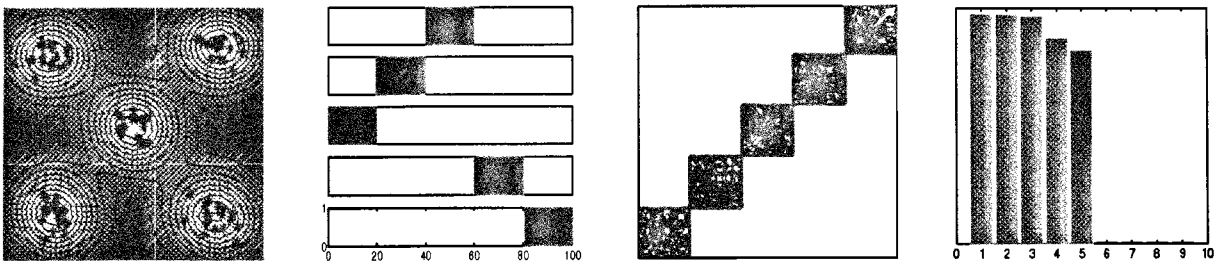
IV. Stochastic Optimization

Stochastic methods for optimizing clustering criteria over a set of binary indicator variables have been suggested in [1] and [3]. In [1], a stochastic method for minimizing the clustering cost based on deterministic annealing was developed. Essentially the cost associated with the overall cluster assignments of the data sample are considered as random variables which have a Gibbs distribution. The expected values, with respect to the Gibbs distribution, of the indicator variables are then estimated in an iterative manner [1]. We define the following feature space cost or distortion $\mathcal{D}_{kj} = 1 - (1/N_k) \sum_l^N z_{kl} K_{jl}$. The term $\mathcal{D}_{kj}$ is the distortion or penalty associated with assigning the jth datum to the kth cluster in feature space. Note that due to the specific use of the RBF kernel the term $(1/N_k) \sum_l^N z_{kl} K_{jl}$ can be viewed as a nonparametric Parzen estimate of the conditional probability of the jth datum given the kth cluster, i.e., $\hat{p}(\mathbf{x}_j|C_k)$. So the penalty or cost associated with assigning the jth datum to the kth cluster in an RBF kernel defined feature space is given as $\mathcal{D}_{kj} = 1 - \hat{p}(\mathbf{x}_j|C_k)$, thus highly improbable points allocated to a cluster will increase the overall clustering cost. Now for an RBF kernel the following holds:

$$\begin{aligned} \mathrm{Tr}\left(\mathbf{S}_W^{\Phi}\right) &= 1 - \frac{1}{N} \sum_{j}^{N} \sum_{k}^{K} z_{kj} \sum_{l}^{N} \frac{z_{kl}}{N_k} K_{jl} \\ &= \frac{1}{N} \sum_{j}^{N} \sum_{k}^{K} z_{kj} \mathcal{D}_{kj} \end{aligned}$$

and minimization of the feature space "sum-of-squares" criterion $\mathrm{Tr}(\mathbf{S}_W^{\Phi})$ corresponds to minimization of $\sum_j^N \sum_k^K z_{kj} \mathcal{D}_{kj}$.

For a general data space sum-of-squares error E_{kn} (a squared distance of datum point n to cluster center k) [1] the overall clustering cost $\mathrm{Tr}(\mathbf{S}_W) = N^{-1} \sum_n^N \sum_k^K z_{kn} E_{kn}$ can be minimized using the



(a) (b) (c) (d)

Fig. 1. (a) The scatter plot of 100 points, composed of 20 datums drawn from five compact and well separated spherical Gaussian clusters. The iso-contours show the lines of constant value of $1 - \mathcal{D}_{kj}$, light colors indicate high values whereas dark colors indicate low values. This was generated using an RBF kernel of width 0.1. (b) This plot shows the value of the binary indicator variables $\mathbf{Z}$ after convergence of the iterative routine to optimize the feature space sum-of-squares clustering criterion. Each row corresponds to a cluster center and the individual data points, ordered in terms of cluster membership (purely for demonstrative purposes) run along the horizontal axis. The bars indicate a value of z_{kj}. It can be seen that there are no cluster assignment errors on this simple data set. (c) The contour plot of the 100×100 kernel matrix clearly showing the inherent block structure. The specific ordering has been used merely for purposes of demonstration and does not affect the results given by the proposed method. (d) The contribution to $\mathbf{1}_N^T \mathbf{K} \mathbf{1}_N$ from the most significant terms $\lambda_i \{\mathbf{1}_N^T \mathbf{u}_i\}^2$. It is most obvious that only five terms contribute to the overall value thus indicating that there are five dominant generators within the data sample.

following iterative procedure, which is reminiscent of an expectation maximization (EM) algorithm

$$\langle z_{kn} \rangle = \frac{\exp(-\beta E_{kn}^{new})}{\sum_{k'=1}^{K} \exp(-\beta E_{k'n}^{new})} \tag{10}$$

and each $E_{kn}^{new} = \|\mathbf{x}_n - \langle \mathbf{m}_k \rangle\|^2$ is re-computed using the new estimates of the expected values of the indicator variables $\langle z_{kn} \rangle$ where

$$\langle \mathbf{m}_k \rangle = \left\{ \sum_{i=1}^{N} \langle z_{ki} \rangle \mathbf{x}_i \right\} \Bigg/ \left\{ \sum_{j=1}^{N} \langle z_{kj} \rangle \right\}.$$

The parameter β controls the *softness* of the assignments during optimization [1]. The reader should refer to [1] and the references therein for a detailed exposition and derivation of the iterative optimization of the central clustering criterion, (2), in data space. This can be straightforwardly used for the proposed feature-space criterion. Employing the distance from the cluster center in feature space defined by (5) and noting that $K_{nn} = 1$ for the RBF kernel some straightforward manipulation yields

$$\langle z_{kn} \rangle = \frac{\exp(-\beta y_{kn}^{new})}{\sum_{k'=1}^{K} \exp(-\beta y_{k'n}^{new})} = \frac{\alpha_k \exp(-2\beta \mathcal{D}_{kn}^{new})}{\sum_{k'=1}^{K} \alpha_{k'} \exp(-2\beta \mathcal{D}_{k'n}^{new})} \tag{11}$$

where

$$\begin{aligned} \alpha_k &= \exp\left\{ -\frac{\beta}{\langle N_k^2 \rangle} \sum_{i=1}^{N} \sum_{j=1}^{N} \langle z_{ki} \rangle \langle z_{kj} \rangle K_{ij} \right\} \\ &= \exp\{-\beta \langle \mathcal{R}(\mathbf{x}|\mathbf{C_k}) \rangle\} \end{aligned}$$

and as such the following iterative procedure (and direct feature space analogs of the data space method) will find a minimum of $\mathrm{Tr}(\mathbf{S}_W^{\Phi})$ [(4)]

$$\langle z_{kn} \rangle = \frac{\alpha_k \exp(-2\beta \mathcal{D}_{kn}^{new})}{\sum_{k'=1}^{K} \alpha_{k'} \exp(-2\beta \mathcal{D}_{k'n}^{new})}$$

and

$$\mathcal{D}_{kn}^{new} = 1 - \frac{1}{\langle N_k \rangle} \sum_{l=1}^{N} \langle z_{kl} \rangle K_{nl}. \tag{12}$$

As the parameter $\beta \to \infty$ then the assignments become hard such that $\langle z_{kn} \rangle \in \{0, 1\}$, i.e., only takes the values zero or one, in which case this becomes the standard batch form of the K-means algorithm in the feature space defined by the RBF kernel $\mathbf{K}$. Note that the term α_k is indicative of the compactness of the kth cluster.

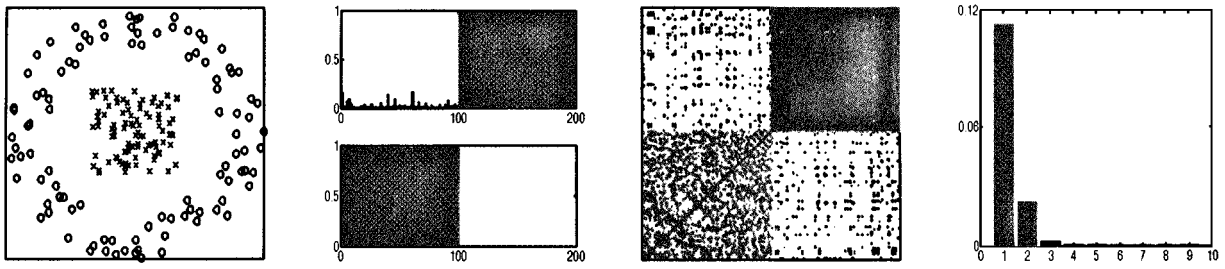
The main point of this proposed method, and indeed most clustering methods is a knowledge of the number of clusters K. The following section proposes a means of estimating the possible number of clusters within the data sample based on the kernel matrix created from the sample of points.

V. Estimating the Number of Clusters Using the Kernel Matrix

Whereas in data space a $D \times N$ dimensional data matrix requires to be manipulated for the optimization of the sum-of-squares criterion, the feature-space counterpart now requires the manipulation of an $N \times N$ dimensional symmetric kernel matrix $\mathbf{K}$. As each element of the kernel matrix defines a dot-product distance in the kernel defined feature space the matrix will have a block diagonal structure when there are definite groupings or clusters within the data sample. This can be clearly seen with a simple example using a two-dimensional (2-D) sample of 100 datum points of which 20 points are each drawn from five spherical Gaussians of variance 0.1 with mean values $\{0, 0; 0.7, 0.7; -0.7, 0.7; 0.7, -0.7; -0.7, -0.7\}$. Fig. 1(a) shows the plot of the data points, the contours show the lines of constant $1 - \mathcal{D}_{kj}$ value for an RBF kernel, i.e., one minus the feature space cost. It is also worth commenting that these contours are also lines of estimated equiprobability. Fig. 1(c) shows the structure of the 100×100 kernel matrix using an RBF kernel of width 0.1. The block structure of the matrix is most apparent. It should be stressed here that the ordering of the points in the figure is purely for illustrative purposes. However, it is to be noted that the eigenvectors of a permuted matrix are the permutations of the original matrix and therefore an indication of the number of clusters within the data may be given from the eigenvalue decomposition of the kernel matrix.

As noted in the previous sections the following finite sample approximation can be made $\int_{\mathbf{x}} p(\mathbf{x})^2 \, d\mathbf{x} \approx (1/N^2) \sum_{i=1}^{N} \sum_{j=1}^{N} K_{ij}$ which can be written in vector/matrix notation as $\mathbf{1}_N^T \mathbf{K} \mathbf{1}_N$ where the $N \times 1$ dimensional vector $\mathbf{1}_N$ has elements of value 1/N. An eigenvalue decomposition on the kernel matrix gives $\mathbf{K} = \mathbf{U} \mathbf{\Lambda} \mathbf{U}^T$ where the columns of the matrix $\mathbf{U}$ are the individual eigenvectors $\mathbf{u}_i$ of $\mathbf{K}$ and the diagonal matrix $\mathbf{\Lambda}$ contains the associated eigenvalues denoted as λ_i. Then we can write

$$\mathbf{1}_N^T \mathbf{K} \mathbf{1}_N = \mathbf{1}_N^T \left\{ \sum_{i=1}^{N} \lambda_i \mathbf{u}_i \mathbf{u}_i^T \right\} \mathbf{1}_N = \sum_{i=1}^{N} \lambda_i \left\{ \mathbf{1}_N^T \mathbf{u}_i \right\}^2. \tag{13}$$



(a) (b) (c) (d)

Fig. 2. (a) The scatter plot of the "Ring Data," 100 samples from a uniform distribution centered at the origin and 100 samples uniformly drawn from an annular ring. (b) The outcome of the clustering method showing that there are no partition errors. (c) The contour plot of the associated kernel matrix (RBF width of 1.0), again note the block diagonal structure. (d) The contribution to $\mathbf{1}_N^T \mathbf{K} \mathbf{1}_N$ from the most significant terms $\lambda_i \{\mathbf{1}_N^T \mathbf{u}_i\}^2$. It is most obvious that only two terms significantly contribute to the overall value thus indicating that there are two dominant generators within the data sample.

The final form of (13) indicates that if there are K distinct clustered regions within the N data samples then there will be K dominant terms $\lambda_i \{\mathbf{1}_N^T \mathbf{u}_i\}^2$ in the summation. Therefore this eigenvalue decomposition method provides a means of estimating the possible number of clusters within the data sample.

It is noted that what has been termed the kernel or Gram matrix [10] in this paper and within the neural computing research community is often referred to as the affinity or proximity matrix within the domain of machine vision research [6]. This affinity matrix is directly analogous to the kernel matrix discussed herein. The segmentation of images into, for example, foreground figures and background is attempted by utilizing the first eigenvector of the affinity/proximity matrix of a particular image [6]. However, no use is made of subsequent eigenvectors in determining the possible number of distinct areas of the image in a manner akin to the cluster number determination method proposed in this letter and so this may indeed be an interesting area of further investigation.

The notion of clustering a data set after it has been nonlinearly transformed into a possibly infinite dimensional feature space has been proposed. A stochastic method for minimizing the trace of the feature space within-group scatter matrix has been suggested. In the case of the feature space whose dot-product is defined by the RBF kernel then a specific form of stochastic iterative update has been developed. The sum-of-squares error in the RBF defined feature space can be viewed as the loss defined by the estimated conditional probability of the datum coming from a particular cluster. The possible number of clusters within the data can be estimated by considering the terms of the eigenvalue decomposition of the kernel matrix created by the data sample. The following section provides some preliminary simulations for demonstrative purposes.

VI. Simulations

To briefly demonstrate the feature space method presented, one toy simulation is given along with some examples provided in [5] and [7]. Fig. 1 shows the results of applying the method to a simple clustered set of data, both the estimation of the number of clusters and the resultant partitioning highlights the effectiveness of this method. Fig. 2 shows the results of applying this method to the 2-D Ring data which originally appeared in [7]. This data is particularly interesting in that the mean (or prototype) vectors in data space for each class coincide. By performing the clustering in a kernel defined feature space the prototypes are therefore calculated in this feature space, which means that they do not necessarily have a pre-image in input space [10]. The implication of this is that the mean vectors in feature space may not serve as representatives or prototypes of the input space clusters. Both the estimation of the number of data generators and the eventual partitioning show the performance of the method on distinctly nonlinearly separable and nonellipsoidal data. These results are identical to the maximum-certainty approach proposed in [7].

Three standard test data sets[1] are employed in the following simulation. The Fisher Iris data is a well-known data collection consisting of four measurements from 50 samples of three varieties of Iris (*Versicolor, Virginica, Setosa*). One of the classes (clusters) is linearly separable from the other two, while the remaining two are not linearly separable. Fig. 3 shows both the clustering achieved and the estimated number of clusters. The number of clusters is estimated correctly and the partition error matches the state of the art results on this data reported in [7], [5]. The next simulation uses the 13-dimensional Wine data set. This data has three classes, varying types of wine, and the 13 features are then used to assign a sample to a particular category of wine. This data has only been investigated in an unsupervised manner in [7] where four partition errors were incurred. Fig. 3 shows the estimated number of data generators using the proposed method. There are only three significant contributors thus indicating the presence of three clusters within the data. Applying the feature space partitioning method yields four errors. The final example is the Crabs data, which consists of five physical measurements of the male and female of two species of crab. Employing the method proposed in this paper correctly estimates the number of possible data clusters.

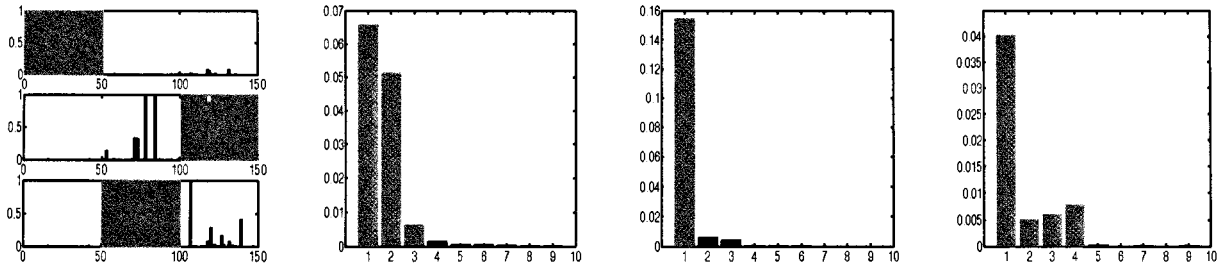
The assessment of the contribution of each term $\lambda_i \{\mathbf{1}_N^T \mathbf{u}_i\}^2$ to the overall value requires some comment. In the case where the clusters in the data are distinct then a pattern similar to that of Figs. 1 and 2 will emerge and the contribution of each term will also be distinct. If, as an example, we consider the Iris data, Fig. 3, it is clear that there are two dominant terms strongly suggestive of the presence of two clusters. However the inclusion of the third smaller term provides 99.76% of the overall value indicating the possible presence of a third and less well defined cluster grouping, as indeed is the case. The assessment of the contribution of each term therefore requires to be considered on a case by case basis.

VII. Conclusion and Discussion

This paper has explored the notion of data clustering in a kernel defined feature space. This follows on from the Support Vector classification methods which employ Mercer kernel representations of feature space dot-products and the unsupervised method for performing feature space principal component analysis (KPCA) [11], [10]. Clustering of data in a feature space has been previously proposed in an earlier unpublished[2] version of [10] where the standard K-means algorithm was presented in kernel space by employing the kernel trick. As the sum-of-squares error criterion for data partitioning can also be posed

[1]Iris, Wine, and Crabs data sets are all available from the UCI machine learning repository.

[2]Available at http://www.kernel-machines.org/.



(a) (b) (c) (d)

Fig. 3. (a) Clustering performance on the Iris data set indicating three partition errors. (b) The contribution to $\mathbf{1}_N^T\mathbf{K}\mathbf{1}_N$ from the most significant terms $\lambda_i\{\mathbf{1}_N^T\mathbf{u}_i\}^2$ for the Iris data. An RBF kernel of width 0.5 was used. The three dominant terms contribute 99.76% of the overall value strongly indicating the existence of two highly dominant and one less dominant data generator, i.e., the existence of three possible clusters. (c) The values of $\lambda_i\{\mathbf{1}_N^T\mathbf{u}_i\}^2$ for the Wine data set (RBF width equals 10). Strongly indicating the presence of only three clusters. (d) The values of $\lambda_i\{\mathbf{1}_N^T\mathbf{u}_i\}^2$ for the Crabs data set (RBF width equals 0.001). Strongly indicating the presence of only four clusters.

in a feature space and as this contains only dot-products of feature vectors a very simple form of feature space clustering criterion arises. We note that the K-means algorithm is the *hard-clustering* limiting case, when $\beta \to \infty$, of the deterministic annealing approach adopted in this paper for optimizing the sum-of-squares clustering criterion.

The reader should note that central clustering by optimization of the sum-of-squares criterion [see (2)] has an intuitive interpretation in that the mean vectors act as representatives of the clusters. However, when performing such clustering in a kernel defined feature space the associated mean vectors may not have a pre-image in the original data space (the ring-data is such an example of this). The implication of this is that the solution may break down, if the estimated centroid is replaced by its nearest data vector.

When specifically considering the RBF kernel then the feature space clustering cost has an interpretation based on nonparametric Parzen window density estimation. It has been proposed that the block-diagonal structure of the kernel matrix be exploited in estimating the number of possible data generators within the sample and the subsequent eigendecompostion of the kernel matrix can indicate the possible number of clusters. Some brief simulations have been provided which indicate the promise of this method of data partitioning and shows that it is comparable with current state-of-the-art partitioning methods [5], [7].

The first point which can be raised regarding the proposed method of data partitioning is with regard to the choice of the type of kernel chosen in defining the nonlinear mapping. This is one of the major questions which is under consideration regarding research being undertaken on support vector and kernel methods. Clearly the choice of kernel will be data specific, however in the specific case of data partitioning then a kernel which will have universal approximation qualities such as the RBF is most appropriate. Indeed this paper has shown that the sum-of-squares criterion in an RBF kernel induced feature space is equivalent to one minus the sum of the estimated conditional probabilities of the data given the clusters. This is an appealing interpretation as the Euclidean metric in data space is now replaced by the probability metric in this specific feature space. So then the specific RBF kernel provides a simple and elegant method of feature space data partitioning based on a sum-of-squares criterion as defined in equation (9). If more general nonlinear mappings are being considered (i.e., ones which do not possess Mercer kernels) then great care must be taken to ensure that the nonlinear transformation chosen does not introduce *structure* which is not intrinsically inherent in the data.

The second point which can be raised about this method is then the choice of the RBF kernel width. This particular concern is pervasive in all methods of unsupervised learning, the selection of an appropriate model parameter, or indeed model, in an unsupervised manner. Clearly cross-validation and leave-one-out techniques are required to estimate the width of the kernel in this method. The maximum certainty approach advocated in [7] requires the fitting of a semiparametric mixture of Gaussians to the data to be clustered, as with the method under consideration the number of Gaussian mixtures requires to be selected *a priori* and heuristics or cross-validation methods require to be employed for this matter.

The complete eigenvalue decomposition of the $N \times N$ kernel matrix scales as $\mathcal{O}(N^3)$ and for a reasonably large dataset this may be prohibitive. However, an iterative method for extracting M eigenvectors from an $N \times N$ dimensional kernel matrix which scales as $\mathcal{O}(MN^2)$ is available [8]. As the number of possible clusters will be small in comparison to the overall size of the data sample then computing the important terms and their percentage contribution to the overall value of $\mathbf{1}_N^T\mathbf{K}\mathbf{1}_N$ is much less costly than the complete decomposition of the kernel matrix.

Once the kernel matrix has been defined then only one nonlinear optimization is required in defining the partitioning. This is in contrast to the method proposed in [7] where each candidate partitioning, the outcome of a nonlinear optimization routine, is used in computing the evidence for the partition based on the data. Therefore at least as many nonlinear optimization routines as there are possible clusters will be required. Only one nonlinear optimization is required in the method proposed in this paper once the probable number of clusters has been selected.

REFERENCES

[1] J. M. Buhmann, "Data clustering and data visualization," in *Learning in Graphical Models*, M. I. Jordon, Ed. Boston, MA: Kluwer, 1998.

[2] J. H. Friedman and J. W. Tukey, "A projection pursuit algorithm for exploratory data analysis," *IEEE Trans. Comput.*, vol. 23, pp. 881–890, 1974.

[3] A. D. Gordon and J. T. Henderson, "An algorithm for Euclidean sum-of-squares classification," *Biometrics*, vol. 33, pp. 355–362, 1977.

[4] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice-Hall, 1988.

[5] T-W. Lee, M. S. Lewicki, and T. S. Sejnowski, "ICA mixture models for unsupervised classification of non-Gaussian sources and automatic context switching in blind signal separation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 22, pp. 1–12, Oct. 2000.

[6] G. L. Scott and H. C. Longuet-Higgins, "Feature grouping by relocalization of eigenvectors of the proximity matrix," in *Proc. British Machine Vision Conf.*, 1990, pp. 103–108.

[7] S. J. Roberts, R. Everson, and I. Rezek, "Maximum certainty data partitioning," *Pattern Recognition*, vol. 33, no. 5, 2000.

[8] R. Rosipal and M. Girolami, "An expectation maximization approach to nonlinear component analysis," *Neural Comput.*, vol. 13, no. 3, pp. 505–510, 2001.

[9] V. Roth and V. Steinhage, "Nonlinear discriminant analysis using kernel functions," in *Advances in Neural Information Processing Systems*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge, MA: MIT Press, 1999, vol. 12, pp. 568–574.

[10] B. Schölkopf, A. Smola, and K. R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Comput.*, vol. 10, no. 5, pp. 1299–1319, 1998.

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