Available online at www.sciencedirect.com
ScienceDirect
Neurocomputing 70 (2007) 2441–2448
NEUROCOMPUTING
www.elsevier.com/locate/neucom

On linear separability of data sets in feature space

Degang Chen[a,*], Qiang He[b], Xizhao Wang[b]

[a]Department of Mathematics and Physics, North China Electric Power University, 102206 Beijing, PR China
[b]Department of Mathematics and Computer Science, Hebei University, Baoding, Hebei 071002, PR China

Received 8 July 2006; received in revised form 29 September 2006; accepted 3 December 2006
Communicated by D. Wang
Available online 9 January 2007

Abstract

In this paper we focus our topic on linear separability of two data sets in feature space, including finite and infinite data sets. We first develop a method to construct a mapping that maps original data set into a high dimensional feature space, on which inner product is defined by a dot product kernel. Our method can also be applied to the Gaussian kernels. Via this mapping, structure of features in the feature space is easily observed, and the linear separability of data sets in feature space could be studied. We obtain that any two finite sets of data with empty overlap in original input space will become linearly separable in an infinite dimensional feature space. For two infinite data sets, we present several sufficient and necessary conditions for their linear separability in feature space. We also obtain a meaningful formula to judge linear separability of two infinite data sets in feature space by information in original input space.

Keywords: Linear separability; SVM; Kernel; Feature space

1. Introduction

Support vector machine (SVM) is a new learning theory presented by Vapnik [11,12]. From pattern recognition viewpoint, it can briefly be stated as follows. When a given sample set K is linearly separable, the separating hyper-plane with maximal margin, the optimal separating hyper-plane, is constructed in original input space. When sample set is linearly non-separating, the input vectors are mapped into a high dimensional feature space through some kernel functions. Then in this high dimensional feature space an optimal separating hyper-plane is constructed. The inner product in high dimensional feature space is just the employed kernel, so complex computing of inner product in high dimensional feature space is avoided. This is one of the advantages of SVM. SVM has been shown to provide higher performance than traditional learning machines [1] and has been introduced as powerful tools for solving classification problems; in the meantime the research on its theory and applications has drawn more and more attention in recent years.

However, if we only consider the computing of inner product in feature space, kernel is enough, it is unnecessary to consider the mapping from original data set to feature space. But if we want to know more about SVM such as analysis of the shape of mapped data in feature space and structure of features in feature space, the mapping from original data set to feature space cannot be ignored. In existing statistical learning theory [7], there are mainly two approaches to obtain the mapping from original data set to feature space. One is to employ the well known Mercer Theorem. By this way the mapping is constructed as a vector whose entries are N_H eigenfunctions of an integral operator, and the kernel corresponds a dot product in $l_2^{N_H}$. Another approach is to consider the Reproducing Kernel Hilbert Space. By this way each pattern is turned to a function on the domain. In this sense, a pattern is now represented by its similarity to all other points in input domain.

However, for the first approach, sometimes it is very difficult to compute the eigenvalues and eigenfunctions of an integral operator defined by a kernel even when they

*Corresponding author. Tel.: +86 13683298633.
E-mail address: chengdegang@263.net (D. Chen).

0925-2312/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2006.12.002

really exist. For the second approach, structures of features are difficult to observe since the image of every input pattern is a function and not a vector. All of these two approaches are mainly designed from the mathematical viewpoint to ensure the existence of such mapping. They are too abstract to be applied to analysis practical problems. Thus an intuitive and general method to construct the mapping from original data set to feature space with legible feature structure is clearly necessary from both theoretical and practical viewpoints.

As is well known, dot product kernels are an important class of kernels in common use. The well known dot product kernels in theory of SVM are homogeneous polynomial kernels, inhomogeneous polynomial kernels. Both homogeneous polynomial kernels and inhomogeneous polynomial kernels map original data set into a finite dimensional polynomial space (feature space) and structures of features are clear (there is a whole field of pattern recognition research studying polynomial classifiers [8]). By using the power series expansion of a dot product kernel, we can develop a mapping from the original data set into a polynomial space (may not be finite dimensional) for every dot product kernel. Via this mapping, structures of features are clear. This method can also be applied to the Gaussian kernels and obtain the corresponding mapping to the feature space.

Furthermore, linear separability of data sets in feature space can also be investigated with this mapping. It can be proven the images of any finite data set are linear independent in the feature space relative to certain dot product kernels, this implies any two finite subclasses of the original data set are linear separable in the feature space. Thus separating hyper-plane with maximal margin, i.e., the optimal separating hyper-plane, is always available. On the other hand, in machine learning society, less effort has been put on linear separability of two infinite data sets. At first glance, it is unnecessary to consider infinite data sets since data sets we deal with in practical problems are all finite. This opinion is from the viewpoint of designing algorithm for practical applications. If we consider the classification problem from theoretical viewpoint, the following three arguments indicate it is meaningful to investigate linear separability of infinite data sets.

First, separating two finite sets linearly is equivalent to separating their convex hulls linearly, and their convex hulls are infinite sets, so we have implicitly considered the linear separability of special infinite data set when separating finite sets linearly. Second, most feature values are real valued, this implies the possible candidate samples may be infinite. So after we construct a learning machine based on finite independent and identically distributed samples, the possible candidate sample we deal with by this machine is always drowned from an infinite set; this also needs to take account of all possible cases drowned according to a probability distribution. As mentioned by Vapnik V. N. [13], the most difficult task is the classification of "candidate boundary points" when all candidate samples are considered, and the SVM technique ignores this difficulty since it just considers finite sample sets. Clearly it does not mean this difficulty does not exist or is solved totally; thus consideration of linear separability of infinite data sets is necessary to develop a method to deal with the candidate boundary points. At last, for a practical binary classification problem, certainly we desire to know the existence and uniqueness of optimal hyper-plane that can separate all candidate samples without misclassification, this also inspires us to consider all candidate samples as an infinite set.

Thus it is necessary to investigate linear separability of infinite data sets at least from the theoretical viewpoint, and such investigation can offer guidance to improve algorithm for practical problems. In this paper we develop several sufficient and necessary conditions for two infinite data sets being linear separable in the feature space, we also obtain a meaningful formula as a sufficient condition to judge linear separability of two infinite data sets in feature space by information of original input space.

This paper is organized as follows. In Section 2 we mainly review some basic content of kernels in SVM. In Section 3 the method of constructing mapping for dot product kernels is developed, and linear separability of two finite data sets is also discussed. In Section 4 we mainly discuss linear separability of two infinite sets in feature space via our proposed mapping.

2. Kernels for SVM

In this paper we only consider binary classification problem. Let $\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset R^n \times \{+1, -1\}$ be a training set, A is the sample set with label $+1$ and B is the sample set with label -1. A and B are called linear separable in R^n if there is a hyper-plane $\langle w, x\rangle + b = 0$ (here $\langle w, x\rangle$ denotes the inner product) and $\delta > 0$ such that $\langle w, x\rangle + b > \delta$ for $x \in A$ and $\langle w, x\rangle + b < -\delta$ for $x \in B$ (this definition is also suitable when A and B are infinite set in a general Hilbert space), clearly $d(A, B) > 0$ holds when A and B are linear separable, and separating hyper-plane with the maximal margin, the optimal separating hyper-plane, could be constructed in R^n. If A and B are not linear separable in R^n, the SVM learning approach projects input patterns x_i with a nonlinear function $\Phi : x \to \Phi(x)$ into a higher dimension space Z and, then, it separates the data in Z with a maximal margin hyper-plane. Therefore, the classifier is given by $f(x) = sign(w^{\mathrm{T}}\Phi(x) + b)$ and parameters w and b are obtained through the minimization of functional $\tau(w) = \frac{1}{2}\|w\|^2$ subject to $y_i(\langle w, x_i\rangle + b) \geqslant 1$ for all $i = 1, \ldots, l$. Since the solution of the linear classifier in Z only involves inner products of vectors $\Phi(x_i)$, we can always use the kernel trick [7], which consists of expressing the inner product in Z as an evaluation of a kernel function in the input space $\langle \Phi(x), \Phi(y)\rangle = k(x, y)$. By this way, we do not need to explicitly know $\Phi(\cdot)$ but just its associated kernel $k(x, y)$. When expressed in terms of kernels, the classifier results in $f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i k(x_i, x) + b)$, where

coefficients $\{\alpha_i\}$ are obtained after a QP optimization of functional $L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{l}\alpha_i\{[\langle x_i,w\rangle-b]y_i-1\}$ which can be solved by the KKT complementarity conditions of optimization theory [1].

However, if we not only consider the computing of inner product in feature space, but also aim at presenting deep insight to SVM such as analysis of the shape of mapped data and structure of features in the feature space, we must deal with the mapping from original data set into the feature space. As pointed in [7], there are mainly two approaches to develop the mapping. One is utilization of the well known Mercer Theorem. Suppose X is a nonempty compact set and $k\in L_\infty(X^2)$ is a kernel, then the integral operator $T_k: L_2(X)\to L_2(X)$ defined as $(T_kf)(x)=\int_\chi k(x,x')f(x')\,d\mu(x')$ is positive definite. Let $\psi_j\in L_2(X)$ be N_H normalized orthogonal eigenfunctions of T_k associated with the eigenvalues $\lambda_j>0$, then $k(x,x')$ corresponds to a dot product in $l_2^{N_H}$ with $\Phi: X\to l_2^{N_H}$ defined as $\Phi(x)=(\sqrt{\lambda_j}\psi_j(x))_{j=1,\ldots,N_H}$. For this method, sometimes it is very difficult to compute the eigenvalues and eigenfunctions of T_k even when they really exist.

Another approach is utilizing the Reproducing Kernel Hilbert Space. We can define a map from X into the space of functions mapping X into R, denoted as $R^X=\{f: X\to R\}$, via $\Phi(x)=k(x',x)$, $x'\in X$, the feature space is spanned by k and is a Reproducing Kernel Hilbert Space. Clearly $\Phi(x)=k(x',x)$ is a function and not a vector, and structures of features are hardly to be observed.

Two kinds of kernels are always applied in SVM [7,10]. They are translation invariant kernels and dot product kernels. The translation invariant kernels are independent of the absolute position of input x and only depend on the difference between two inputs x and x', so it can be denoted as $k(x,x')=k(x-x')$. The well known translation invariant kernel is the Gaussian radial basis function kernel $k(x,x')=\exp(-\|x-x'\|^2/2\sigma^2)$, other translation invariant kernels include B_n—splines kernels [9], Dirichlet kernels [7] and Periodic kernels [7]. A second important family of kernels can be efficiently described in terms of dot product, i.e., $k(x,x')=k(\langle x,x'\rangle)$. The well known dot product kernels are Homogeneous Polynomial Kernels $k(x,x')=\langle x,x'\rangle^p$, inhomogeneous Polynomial Kernels $k(x,x')=(\langle x,x'\rangle+c)^p$ with $c\geqslant 0$. Both Homogeneous Polynomial Kernels and inhomogeneous Polynomial Kernels map input set into a finite dimensional Polynomial space. In [2] we have also considered a class of infinite Polynomial kernels on a compact subset U_n of the open unit ball $\{x\in R^n:\|x\|<1\}$, defined as $k_c(x,x e)=(1-\langle x,x'\rangle)^p/(1-\langle x,x'\rangle)^p$, for every $x,x'\in U_n$, $p\in N-\{1\}$, via an infinite Polynomial kernel, the input data set is projected into an infinite dimensional Polynomial space.

3. The mapping for dot product kernels

In this section we will focus on developing a general method to construct the mapping from the original data set into the feature space for the dot product kernels. This method is also suitable to deal with the Gaussian kernels. We can prove if the feature space is an infinite dimensional Polynomial space, then any two finite sets of data in original space will become linearly separable in feature space.

For dot product kernels, the following theorem proposed in [6] is always useful.

Theorem 1 (*Schoenberg [6]*). *A function* $k(x,x')=k(\langle x,x'\rangle)$ *defined on an infinite dimensional Hilbert space, with a power series expansion* $k(t)=\sum_{n=0}^{\infty}a_nt^n$ *is a positive definite kernel if and only if for all* n, *we have* $a_n\geqslant 0$.

This theorem implies many kinds of dot product kernels that can be considered in SVM.

Suppose $k(x,x')=k(\langle x,x'\rangle)$ is a dot product kernel on $X\subset R^n$ with power series expansion:

$$k(\langle x,x'\rangle)=\sum_{n=0}^{\infty}a_n\langle x,x'\rangle^n. \tag{1}$$

For every $x\in X$, define C_n to map $x\in X$ to vector $C_n(x)$ whose entries are all possible nth degree ordered products of the entries of x, and define Φ_k by compensating for the multiple occurrence of certain monomials in C_n by scaling the respective entries of Φ_n with the square roots of their numbers of occurrence. Then, by the construction of C_n and Φ_n, we have

$$\langle C_n(x),C_n(x')\rangle=\langle\Phi_n(x),\Phi_n(x')\rangle=\langle x,x'\rangle^n. \tag{2}$$

This fact can be found in [7] and is well known for the Homogeneous Polynomial Kernels $k(x,x')=\langle x,x'\rangle^p$.

Define $\Phi(x)=(a_0,\sqrt{a_1}\Phi_1(x),\ldots,\sqrt{a_n}\Phi_n(x),\ldots,)$, then we have

$$\langle\Phi(x),\Phi(x')\rangle=k(x,x'). \tag{3}$$

Clearly $\Phi_1(x)=x$ holds, this implies if $a_1\neq 0$, then $\Phi(x)$ is the extension of x by adding features and keeps all the original entries of x, thus $\Phi(x)$ keeps the original information of x. This statement is a goodness of Φ. Entries of $\Phi(x)$ are constructed by entries of x, thus structure of appending features are clear and easy to be analyzed since these appending features are constructed by the original features. The feature space with respect to $k(x,x')$ can be selected as the Hilbert space spanned by $\Phi(X)$ and is denoted by $H^k(X)$.

First we consider the properties of the above proposed Φ when the feature space is finite dimensional. If there is $n_0\in N$ such that $a_n=0$ when $n>n_0$, then we have

$$k(x,x')=\sum_{n=0}^{n_0}a_n\langle x,x'\rangle^n, \tag{4}$$

thus $k(x,x')$ is just the weighted sum of some Homogeneous Polynomial Kernels, and the feature space is a finite dimensional Homogeneous Polynomial space. However, for $k(x,x')=\langle x,x'\rangle^n$, it is possible that Φ is not a one to one mapping, i.e., different inputs may have the same

image, which is clearly unreasonable. This statement can be illustrated by the following example.

Example 1. If $n=2$, and $x=(x_1,\ x_2)$, then $\Phi(x)=\Phi_2(x)=(x_1^2,x_2^2,\sqrt{2}x_1x_2)$. For two different inputs $x=(1,\ -1)$, $y=(-1,1)$, clearly $x\neq y$, but $\Phi(x)=\Phi(y)$. If x and y belong to different classes, then every separating hyper-plane in feature space relative to the kernel $k(x,x')=\langle x,x'\rangle^2$ cannot distinguish x and y. Similar cases will appear frequently when n is even. If we select a weighted sum form kernel predigest satisfying $a_1\neq 0$, then entries of x are a part of $\Phi(x)$, thus we can avoid this case.

By using our proposed Φ, we have the following useful theorem.

Theorem 2. *Suppose* $\{x_1,\ldots,x_m\}\subset X$ *satisfying* $x_i\neq 0$, $x_i\neq x_j$ *if* $i\neq j$, *then there is a dot product kernel*:

$$k(x,x')=\sum_{n=0}^{n_0}a_n\langle x,x'\rangle^n, \tag{5}$$

such that $\Phi(x_1),\ldots,\Phi(x_n)$ *are linear independent*.

Proof. Suppose $x_i=(a_{i1},a_{i2},\ldots,a_{in})$, $k(x,x')=\sum_{n=0}^{m-1}\langle x,x'\rangle^n$, then $k(x,x')$ is a dot product with expression:

$$k(x,x')=\frac{1-\langle x,x'\rangle^m}{1-\langle x,x'\rangle}. \tag{6}$$

Let $f_i(x)=a_{i1}+a_{i2}x+\cdots+a_{in}x^{n-1}$, $i=1,\ldots,m$. If $i\neq j$, then $x_i\neq x_j$; we have that $f_i(x)$ and $f_j(x)$ are two different equations. By the algebraic basic theorem we know every $f_i(x)-f_j(x)=0$ has finite number of roots. Thus there exists $n_0\in N$such that any two of $\{f_i(n_0):i=1,\ldots,m\}$ are different. Let

$$\beta_i=\{1,f_i(n_0),\ldots,f_i^{m-1}(n_0)\},\quad i=1,\ldots,m, \tag{7}$$

then we have that $\beta_1,\beta_2,\ldots,\beta_m$ are linear independent.

Suppose

$$\alpha_1\Phi(x_1)+\alpha_2\Phi(x_2)+\cdots+\alpha_m\Phi(x_m)=0, \tag{8}$$

then for $l_1+l_2+\cdots+l_n\leqslant m-1$, $l_1,l_2,\ldots,l_n\in N\cup\{0\}$, we have

$$\sum_{i=1}^{m}\alpha_i a_{i1}^{l_1}a_{i2}^{l_2}\ldots a_{in}^{l_n}=0, \tag{9}$$

we have $\sum_{i=1}^{m}\alpha_i f_i^n(n_0)=0$, this implies $\alpha_1\beta_1+\cdots+\alpha_m\beta_m=0$, thus every $\alpha_i=0$ and $\Phi(x_1),\ldots,\Phi(x_n)$ are linear independent. □

In proof of Theorem 2 we choose the kernel as $k(x,x')=(1-\langle x,x'\rangle)^m/(1-\langle x,x'\rangle)$ in order to predigest the proof. However, every kernel $k(x,x')=\sum_{n=0}^{n_0}a_n\langle x,x'\rangle^n$ satisfying $n_0\geqslant m-1$ and $a_n>0$ for $n\leqslant m-1$ satisfies the condition in Theorem 2.

Suppose Φ is a mapping relative to a kernel $k(x,x')$ such that $\Phi(x_1),\ldots,\Phi(x_n)$ are linear independent, A and B are two nonempty subsets of X and $A\cap B=\emptyset$, then we have $\Phi(X)=\Phi(A)\cup\Phi(B)$ and $\Phi(A)\cap\Phi(B)=\emptyset$. That $\Phi(x_1),\ldots,\Phi(x_n)$ are linear independent implies any element in the convex hull of one class cannot be the convex combination of the elements of another class; this implies convex hulls of A and B have empty overlap; notice these two convex hulls are compact, so $\{\Phi(x_1),\ldots,\Phi(x_l)\}$ and $\{\Phi(x_{l+1}),\ldots,\Phi(x_m)\}$ are linear separable in feature space. Thus we can derive the following theorem.

Theorem 3. *Suppose* $\{(x_1,y_1),\ldots,(x_l,y_l)\}\subset X\times\{+1\}$, $\{(x_{l+1},y_{l+1}),\ldots,(x_m,y_m)\}\subset X\times\{-1\}$, *then there is a mapping relative to a dot product kernel which map X into a finite dimensional Polynomial space such that these two classes are linear separable in feature space*.

Proof. It is straightforward according to Theorem 2 and the discussion in the previous paragraph. □

Suppose $k(\langle x,x'\rangle)=\sum_{n=0}^{\infty}a_n\langle x,x'\rangle^n$ satisfies for every $n_0\in N$ there exists $n>n_0$ such that $a_n>0$, without losing universality, we assume every $a_n>0$, i.e., every coefficient in its power series is positive, for example, Vovk's infinite polynomial kernel $k(x,x')=(1-(\langle x,x'\rangle))^{-1}$ [5,7] and our proposed infinite polynomial kernel $k_c(x,x')=(1-\langle x,x'\rangle)^p/(1-\langle x,x'\rangle)^p$ [2]. The following theorem implies the feature space relative to such kernels is infinite dimensional.

Theorem 4. *Suppose* $\{x_1,\ldots,x_m\}\subset X$ *satisfies* $x_i\neq 0$ *for* $i=1,2,\ldots,m$, $x_i\neq x_j$ *if* $i\neq j$, Φ *is the mapping relative to* $k(\langle x,x'\rangle)=\sum_{n=0}^{\infty}a_n\langle x,x'\rangle^n$ *such that every* $a_n>0$, *then* $\Phi(x_1),\ldots,\Phi(x_n)$ *are linear independent*.

Proof. Suppose $x_i=(a_{i1},a_{i2},\ldots,a_{in})$ and $\Phi(x_1),\ldots,\Phi(x_m)$ are linear dependent, then there exists $\alpha_1,\alpha_2,\ldots,\alpha_m$ satisfying that at least one of them is not equal to zero and $\alpha_1\Phi(x_1)+\alpha_2\Phi(x_2)+\cdots+\alpha_m\Phi(x_m)=0$ holds. Thus we have $\sum_{i=1}^{m}\alpha_i a_{i1}^{l_1}a_{i2}^{l_2}\ldots a_{in}^{l_n}=0$ where $l_1,l_2,\ldots,l_n\in N\cup\{0\}$.

Let

$$f_i(x)=a_{i1}+a_{i2}x+\cdots+a_{in}x^{n-1},\quad i=1,\ldots,m. \tag{10}$$

Then there exists $n_0\in N$ such that any two of $\{f_i(n_0):i=1,\ldots,m\}$ are different. Let

$$\beta_i=\{1,f_i(n_0),\ldots,f_i^{m-1}(n_0)\},\quad i=1,\ldots,m, \tag{11}$$

then we have $\beta_1,\beta_2,..,\beta_m$ are linear independent. But by $\sum_{i=1}^{m}\alpha_i a_{i1}^{l_1}a_{i2}^{l_2}\ldots a_{in}^{l_n}=0$ we have

$$\alpha_1\beta_1+\cdots+\alpha_m\beta_m=0, \tag{12}$$

this is a contradiction. Thus we have $\Phi(x_1),\ldots,\Phi(x_n)$ are linear independent. □

For $\{x_1,\ldots,x_m\}\subset X$, Theorem 2 implies there exists a finite dimensional feature space such that images of $\{x_1,\ldots,x_m\}$ are linear independent in this feature space, while Theorem 4 implies images of $\{x_1,\ldots,x_m\}$ are linear independent in the feature space relative to a kernel satisfying that every coefficient in its power series is positive, so these two theorems are different. For the

kernel satisfying that every coefficient in its power series is positive, similar to Theorem 3 we have the following result.

Theorem 5. *Suppose* $\{(x_1, y_1), \ldots, (x_l, y_l)\} \subset X \times \{+1\}$, $\{(x_{l+1}, y_{l+1}), \ldots, (x_m, y_m)\} \subset X \times \{-1\}$, *then they are linear separable in every feature space relative to a kernel satisfying that every coefficient in its power series is positive.*

However, as pointed in Section 2, for a fixed kernel $k(x, x')$, the feature space is not uniqueness. The following theorem implies selection of feature space does not influence the linear independence of a finite class of data in feature space.

Theorem 6. *Suppose* $\{x_1, \ldots, x_m\} \subset X$ *satisfies* $x_i \neq 0$ *for* $i = 1, 2, \ldots, m$, $x_i \neq x_j$ *if* $i \neq j$, *then the Gram matrix* $M = (k(x_i, x_j))$ *has full rank for a dot product kernel* $k(x, x')$ *satisfying that every coefficient in its power series is positive.*

Proof. If $M = (k(x_i, x_j)) = (\langle \Phi(x_i), \Phi(x_j) \rangle)$ has no full rank, then there exists $\alpha_1, \alpha_2, \ldots, \alpha_m$ satisfying that at least one of them is not equal to zero such that

$$\sum_{l=1}^{m} \alpha_l \langle \Phi(x_l), \Phi(x_i) \rangle = 0, \quad i = 1, \ldots, m. \tag{13}$$

So we have

$$\left\langle \alpha_i \Phi(x_i), \sum_{l=1}^{m} \alpha_l \Phi(x_l) \right\rangle = 0, \quad i = 1, \ldots, m \tag{14}$$

which implies

$$\left\langle \sum_{i=1}^{m} \alpha_i \Phi(x_i), \sum_{l=1}^{m} \alpha_l \Phi(x_l) \right\rangle = 0, \tag{15}$$

thus $\sum_{i=1}^{m} \alpha_i \Phi(x_i) = 0$ and $\Phi(x_1), \ldots, \Phi(x_n)$ are linear dependent. Hence $M = (k\langle x_i, x_j \rangle) = \langle \Phi(x_i), \Phi(x_j) \rangle$ has full rank.

If Φ' is another mapping that projects X into a different feature space, then it is easy to prove $\Phi'(x_1)$, $\Phi'(x_2), \ldots, \Phi'(x_m)$ are linear independent by $M = (k(x_i, x_j))$ has full rank. □

For two dot product kernels k_1 and k_2, suppose Φ_1 and Φ_2 are mappings relative to k_1 and k_2, respectively, we have the following straightforward but useful theorem.

Theorem 7. *If* Φ_2 *is the extension of* Φ_1, *then* $\Phi_1(x_1), \ldots, \Phi_1(x_n)$ *that are linear independent implies* $\Phi_2(x_1), \ldots, \Phi_2(x_n)$ *are linear independent.*

Our proposed method to construct mapping for dot product kernels can be applied to the Gaussian kernels on the surface of the unit ball. Suppose every $x \in X$ is a unit vector, i.e., $\|x\| = 1$, then $\|x - x'\|^2 = \langle x - x', x - x' \rangle = 2 - 2\langle x, x' \rangle$, thus the Gaussian kernels $k(x, x') = \exp(-\|x - x'\|^2/2\sigma^2)$ have an equivalence expression as dot product kernels as $k(x, x') = \exp((\langle x, x' \rangle - 1)/\sigma^2)$, and we can construct the mapping for the Gaussian kernels by its power series by our proposed method. In [7] it has been pointed the Gaussian Gram Matrices are full rank, i.e., if Φ_G is the mapping relative to a Gaussian kernel, then $\Phi_G(x_1), \ldots, \Phi_G(x_m)$ are linear dependent for $\{x_1, \ldots, x_m\} \subset X$, this statement is very important for analysis of the properties of Gaussian kernels. By Theorem 4 we can also get this conclusion and we propose a new straight proof for this result, our proof is different from the original one in [4].

For a finite data set $\{x_1, \ldots, x_m\} \subset X$, $\Phi(x_1), \ldots, \Phi(x_n)$ are linear independent that implies any binary partition of $\{x_1, \ldots, x_m\}$ are linear separable in the feature space. So $\Phi(x_1), \ldots, \Phi(x_n)$ being linear independent is a sufficient condition of $\{x_1, \ldots, x_m\}$ being linear separable in feature space and clearly not a necessary condition. This sufficient condition illustrates the rationale of the kernel trick in SVM. However, it seems this sufficient condition is too strong since we always just need to separate two subsets of $\{x_1, \ldots, x_m\}$ instead of separating all its possible binary partitions.

4. On linear separability of infinite data sets in feature space

In this section we mainly discuss linear separability of infinite data sets in feature space, while we have emphasized the importance of such a discussion in the introduction. At first glance, it seems that we can employ the well known separation theorems for convex sets in classical functional analysis to solve this problem, but it is really possible that these famous theorems would not work in an infinite dimensional feature space. This can be due to the following reason. As is well known, to separate two convex sets by a hyper-plane with the separation theorems, one necessary condition is that at least one of these two convex sets has nonempty interior, and this condition does not always hold in infinite dimensional feature space since it is spanned by the linear combinations of elements in the given input data sets. Here we do not want to explain this statement in detail since it is just a mathematical problem and is well known in functional analysis. In this section we develop several sufficient and necessary conditions to characterize the linear separability of two infinite sets in an infinite dimensional feature space. We also obtain a formula to judge the linear separability of two infinite sets in the infinite dimensional feature space related to the Gaussian kernels by information of the original input data set.

In this section we consider linear separability of two infinite subsets of surface of the unit ball, i.e., every sample is a unit vector. Suppose $X \subset R^n$ is a compact subset of surface of the unit ball, and we select the infinite dimensional feature space related to the Gaussian kernel $k(x, x') = \exp((\langle x, x' \rangle - 1)/\sigma^2) = \sum_{t=0}^{\infty} a_t \langle x, x' \rangle^t$, denoted as $H^G(X)$. Let $k_n(x, x') = \sum_{t=1}^{n} a_t \langle x, x' \rangle^t$, clearly $k_n(x, x') \to k(x, x')$ uniformly on a compact set, and we denote $\Phi_n : X \to H^{k_n}(X)$ which maps X into the finite dimensional feature space $H^{k_n}(X)$ relative to kernel $k_n(x, x')$.

Theorem 8. *Suppose $X \subset R^n$ is a compact subset of the surface of the unit ball and $X = A \cup B$, $A \cap B = \emptyset$. Then $\Phi(A)$ and $\Phi(B)$ are linear separable in $H^G(X)$ if and only if the crowed point sets of A and B have empty overlap, i.e., the boundary point set of A and B is empty.*

Proof. $\Rightarrow$ Since X is compact, we know crowed points of X are still in X, thus crowed points of $\Phi(X)$ are still in $\Phi(X)$ by $\Phi(X)$ is compact. If crowed point sets of A and B have nonempty overlap, then the crowed point sets of $\Phi(A)$ and $\Phi(B)$ also have nonempty overlap by Φ is continuous, this implies $d(\Phi(A), \Phi(B)) = 0$, so $\Phi(A)$ and $\Phi(B)$ cannot be linear separable in feature space.

$\Leftarrow$ Suppose the crowed point sets of A and B have empty overlap. Clearly A and B are compact, this implies $\Phi(A)$ and $\Phi(B)$ are compact in the feature space in case of Φ being continuous. By Theorem 5 the overlap of convex hulls of $\Phi(A)$ and $\Phi(B)$ that are empty, by the definition of Φ we have convex hulls of $\Phi(A)$ and $\Phi(B)$ that are compact, thus they are linear separable in the feature space and $\Phi(A)$ and $\Phi(B)$ are linear separable in the feature space. □

For a binary pattern recognition problem, if there is a hyper-plane which cannot only separate the training sample but can also classify every possible candidate data properly, i.e., it can separate all the possible data of two classes without misclassification; we call this binary pattern recognition problem can be totally solved. Theorem 8 develops a sufficient and necessary condition under which it is possible to solve totally a binary pattern recognition problem, i.e., for every sample of one class, there exists a sufficient small neighborhood of this sample satisfying that no sample of another class is in this neighborhood. This statement implies the candidate boundary points do not exist for a binary pattern recognition problem which can be totally solved. Thus the problem mentioned by Vapnik V. N. [13] that the most difficult task is the classification of "candidate boundary points" when all candidate samples are considered does not hold for a binary pattern recognition problem which can be totally solved. Also by Theorem 8, if a binary pattern recognition problem can be totally solved, then the distance between the two convex hulls of two different sample classes is strictly bigger than zero. If a binary pattern recognition problem cannot be totally solved, then the distance between the two convex hulls of two different sample classes is equal to zero, i.e., there are candidate boundary points. Thus we can conclude that for a binary pattern recognition problem, if it can be solved totally, then generally the selection of optimal separating hyper-plane is not unique, if it cannot be solved totally, then the optimal separating hyper-plane does not exist. The Fig. 1 illustrates our idea of Theorem 8.

As pointed out in [3], since one has to make assumptions about the structure of data set (otherwise no generalization is possible), it is natural to assume that two points that are close are likely to belong to the same class; informally, we want similar inputs to lead to similar outputs [7]. Most classical classification algorithms rely, implicitly or explicitly, on such an assumption (e.g. nearest-neighbor classifiers, and the simplest possible justification for large margins in SVM in [7]). Applying this assumption to the binary pattern recognition problems, it just implies the crowed points of the two classes have an empty overlap which is the condition in Theorem 8, thus optimal separating hyper-plane in feature space always exists and is not unique.

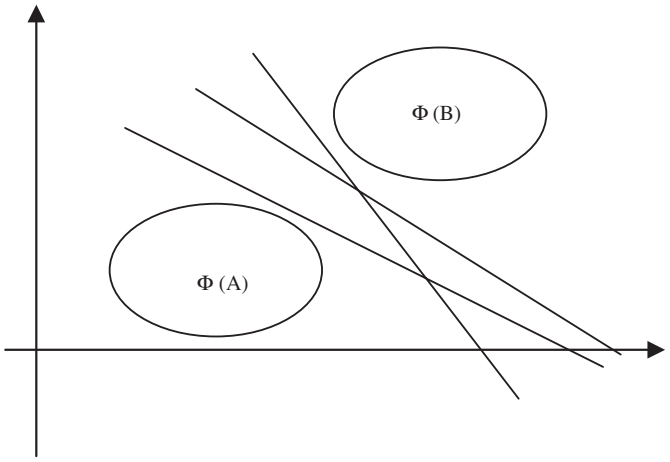

Fig. 1. If X is compact and $X = A \cup B, A \cap B = \emptyset$, all the possible data are in X. If the crowed point sets of A and B have empty overlap, then $\Phi(A)$ and $\Phi(B)$ are linear separable in the feature space as shown in the above figure. Since every separating hyper-plane can classify all the possible input data without misclassification as the three lines in the figure, each of them can be selected as an optimal separating hyper-plane.

If the binary pattern recognition problems do not satisfy this assumption, i.e., the two classes have conjunct crowed points, then the optimal separating hyper-plane that can separate all the data without misclassification is not available. In this way, in an infinite dimensional feature space relative to a dot product kernel, two classes of data distribute along the different sides of the crowed points, and the best separating hyper-plane should pass through the crowed points. As mentioned by Vapnik V. N. [13], the most difficult task is the classification of "candidate boundary points" when all candidate samples are considered, and we propose a reasonable idea to deal with the candidate boundary points. We employ the following simple example to illustrate our idea.

Example 2. Suppose we have two tangent ellipses as two classes, thus the tangent point is the conjunct crowed point. If we want to separate them by a line, then clearly the tangent is the best selection. The following figure (Fig. 2 can explain this example straightforwardly.

Theorem 9. *Suppose $X \subset R^n$ is a compact subset of surface of the unit ball and $X = A \cup B$, $A \cap B = \emptyset$. Then $\Phi(A)$ and $\Phi(B)$ are linear separable in $H^G(X)$ if and only if there exists n_0 such that $\Phi_{n_0}(A)$ and $\Phi_{n_0}(B)$ are linear separable in $H^{k_{n_0}}(X)$, i.e., in the feature space relative to $k_{n_0}(x, x')$.*

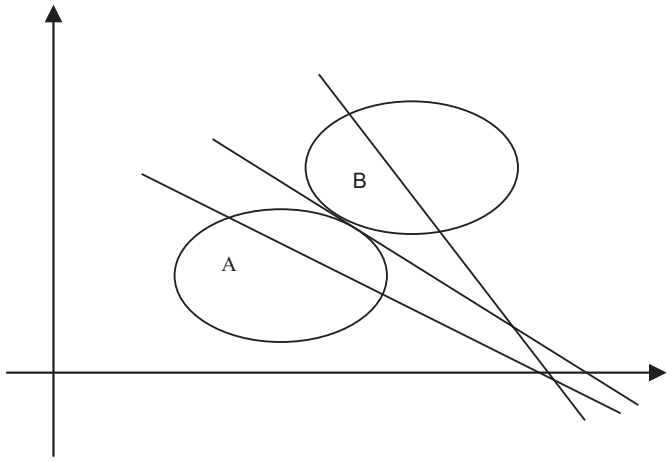

Fig. 2. To separate two tangent ellipses by a line, clearly the tangent is the best selection.

Proof. Here we denote the convex hull of $\Phi(A)$ as $\mathrm{co}(\Phi(A))$. Clearly $\Phi(A)$ and $\Phi(B)$ are linear separable in $H^{\mathrm{G}}(X)$ if and only if $d(\mathrm{co}(\Phi(A)),\mathrm{co}(\Phi(B)))>0\cdot d(\mathrm{co}(\Phi_n(A)),\mathrm{co}(\Phi_n(B)))\to d(\mathrm{co}(\Phi(A)),\mathrm{co}(\Phi(B)))$ increasingly since $k_n(x,x')\to k(x,x')$ uniformly; thus we can get the proof. □

According to Theorem 8, if we assume that two close points are likely to belong to the same class, then two disjoint data sets are linear separable in an infinite dimensional feature space, then by Theorem 9, they must be linear separable in a finite dimensional feature space, and in a finite dimensional feature space the VC dimension of the set of linear classifier is finite which implies better generalized ability [12].

However, Theorems 8 and 9 characterize the linear separability of two data sets from the theoretical viewpoints, It is difficult to apply them to practical problems. We develop a sufficient condition to judge linear separability of two data sets as follows.

Theorem 10. *Suppose $X\subset R^n$ is a compact subset of the surface of the unit ball and $X=A\cup B$, $A\cap B=\emptyset$.* Let

$$d_A=\sup_{x,y\in A}\|x-y\|, \tag{16}$$

$$d_B=\sup_{x,y\in B}\|x-y\|, \tag{17}$$

$$d_{AB}=\inf_{x\in A,y\in B}\|x-y\|. \tag{18}$$

If $\mathrm{e}^{-(d_A)^2/2\sigma^2}+\mathrm{e}^{-(d_B)^2/2\sigma^2}>2\mathrm{e}^{-(d_{AB})^2/2\sigma^2}$ holds, then $\Phi(A)$ and $\Phi(B)$ are linear separable in $H^{\mathrm{G}}(X)$. Here sup *means the supremum and* inf *means the infimum.*

Proof. For $\sum_{i=1}^m\alpha_i\Phi(x_i)\in\mathrm{co}(\Phi(A))$, $\sum_{j=1}^n\beta_i\Phi(y_i)\in\mathrm{co}(\Phi(B))$, here $x_i\in A$, $y_j\in B$, $\sum_{i=1}^m\alpha_i=1$, $\sum_{j=1}^n\beta_j=1$, $\alpha_i\geqslant 0$, $\beta_j\geqslant 0$, $i=1,\ldots,m, i=1,\ldots,n$, We have

$$\begin{aligned}
&\left\|\sum_{i=1}^m\alpha_i\Phi(x_i)-\sum_{j=1}^n\beta_j\Phi(y_j)\right\|^2\\
&=\left\langle\sum_{i=1}^m\alpha_i\Phi(x_i)-\sum_{j=1}^n\beta_j\Phi(y_j),\sum_{i=1}^m\alpha_j\Phi(x_j)-\sum_{j=1}^n\beta_j\Phi(y_j)\right\rangle\\
&=\left\langle\sum_{i=1}^m\alpha_i\Phi(x_i),\sum_{i=1}^m\alpha_i\Phi(x_i)\right\rangle\\
&\quad+\left\langle\sum_{j=1}^n\beta_j\Phi(y_j),\sum_{i=1}^m\beta_j\Phi(y_j)\right\rangle\\
&\quad-2\left\langle\sum_{i=1}^m\alpha_i\Phi(x_i),\sum_{j=1}^n\beta_j\Phi(y_j)\right\rangle.
\end{aligned}\tag{19}$$

And we have

$$\begin{aligned}
&\left\langle\sum_{i=1}^m\alpha_i\Phi(x_i),\sum_{i=1}^m\alpha_i\Phi(x_i)\right\rangle\\
&=\sum_{i=1}^m\sum_{k=1}^m\alpha_i\alpha_k\langle\Phi(x_i),\Phi(x_k)\rangle\\
&=1-\sum_{i\neq k}\alpha_i\alpha_k+\sum_{i\neq k}\alpha_i\alpha_k\mathrm{e}^{-\|x_i-x_k\|^2/2\sigma^2}\\
&\geqslant 1-\frac{m(m-1)}{m^2}(1-\mathrm{e}^{-(d_A)^2/2\sigma^2})\geqslant\mathrm{e}^{-(d_A)^2/2\sigma^2}.
\end{aligned}\tag{20}$$

Similarly we have

$$\begin{aligned}
&\left\langle\sum_{j=1}^n\beta_j\Phi(y_j),\sum_{i=1}^m\beta_j\Phi(y_j)\right\rangle\geqslant\mathrm{e}^{-(d_B)^2/2\sigma^2},\\
&\left\langle\sum_{i=1}^m\alpha_i\Phi(x_i),\sum_{j=1}^n\beta_j\Phi(y_j)\right\rangle\leqslant\mathrm{e}^{-(d_{AB})^2/2\sigma^2}.
\end{aligned}\tag{21}$$

Thus we have

$$\begin{aligned}
&\left\|\sum_{i=1}^m\alpha_i\Phi(x_i)-\sum_{j=1}^n\beta_j\Phi(y_j)\right\|^2\\
&\geqslant\mathrm{e}^{-(d_A)^2/2\sigma^2}+\mathrm{e}^{-(d_B)^2/2\sigma^2}-2\mathrm{e}^{-(d_{AB})^2/2\sigma^2}.
\end{aligned}\tag{22}$$

If $\mathrm{e}^{-(d_A)^2/2\sigma^2}+\mathrm{e}^{-(d_B)^2/2\sigma^2}>2\mathrm{e}^{-(d_{AB})^2/2\sigma^2}$, then $\mathrm{e}^{-(d_A)^2/2\sigma^2}+\mathrm{e}^{-(d_B)^2/2\sigma^2}-2\mathrm{e}^{-(d_{AB})^2/2\sigma^2}=\delta>0$, hence we have

$$\left\|\sum_{i=1}^m\alpha_i\Phi(x_i)-\sum_{j=1}^n\beta_j\Phi(y_j)\right\|^2\geqslant\delta>0, \tag{23}$$

this implies $d(\mathrm{co}(\Phi(A)),\mathrm{co}(\Phi(B)))>0$ and $\Phi(A)$ and $\Phi(B)$ are linear separable in $H^{\mathrm{G}}(X)$. □

The formula $\mathrm{e}^{-(d_A)^2/2\sigma^2}+\mathrm{e}^{-(d_B)^2/2\sigma^2}>2\mathrm{e}^{-(d_{AB})^2/2\sigma^2}$ is meaningful. Since d_A is the diameter of A, d_B is the diameter of B, and d_{AB} is the distance between A and B, all of d_A, d_B and d_{AB} are basic information of the input data set; thus Theorem 10 develops a sufficient condition to characterize the linear separability of two data sets by the

information of original input data set. We can easily have the following result.

Proposition 1. *Suppose* $X \subset R^n$ *is a compact subset of the surface of the unit ball and* $X = A \cup B$, $A \cap B = \emptyset$. *Let* $d_A = \sup_{x,y\in A}\|x-y\|$, $d_B = \sup_{x,y\in B}\|x-y\|$, $d_{AB} = \inf_{x\in A,y\in B}\|x-y\|$. *If* $\max\{d_A, d_B\} < d_{AB}$ *holds, then* $\Phi(A)$ *and* $\Phi(B)$ *are linear separable in* $H^G(X)$.

Notice the distance between two samples in the input space characterizes their similarity; we can conclude with Proposition 1 that if the distance between any two samples in the same class is strictly smaller than the distance between any two samples belonging to different classes (samples belonging to the same class are more similar than samples belonging to different classes), then these two classes must be linear separable in the feature space $H^G(X)$.

5. Conclusion

In this paper we mainly discuss linear separability of two data sets in feature space relative to a dot product kernel. For two finite data sets, we prove that they must be linear separable in an infinite dimensional feature space. For two infinite data sets, we develop two sufficient and necessary conditions to characterize their linear separability in feature space relative to the Gaussian kernel. For a binary classification problem, if we assume that two points that are close are likely to belong to the same class, then these two classes are linear separable in the feature space. We also develop a meaningful sufficient condition to judge the linear separability of two infinite data sets in the feature space by the information of the original input data space. We hope results in this paper could be applied to develop algorithms for pattern recognition with better performance, and this will be our future work.

Acknowledgments

This paper is supported by the Foundation of North China Electric Power University, the National Natural Science Foundation of China(NSFC60473045).

References

[1] C. Burges, A Tutorial on Support Vector Machines for Pattern Recognition, Data Min. Knowl. Discovery 2 (2) (1998) 121–167.
[2] D. Chen, Q. He, X. Wang, The infinite polynomial kernel for support vector machine, Lect. Notes Artif. Intell. 3584 (2005) 267–275.
[3] H. Matthias, B. Olivier, S. Bernhard, Maximal margin classification for metric spaces, J. Comput. Sys. Sci. 71 (2005) 333–359.
[4] C.A. Micchelli, Algebraic aspects of interpolation, in: Proceedings of Symposia in Applied Mathematics, vol. 36, 1986, pp. 81–102.
[5] C. Saunders, M.O. Stitson, J. Weston, L. Bottou, B. Scholkopf, A.J. Smola, Support vector machine reference manual, Technical Report CSD-TR-98-03, Department of Computer Science, Royal Holloway, University of London, Egham, UK, 1998.
[6] I.J. Schoenberg, Positive definite functions on spheres, Duke Math. J. 9 (1942) 96–108.
[7] B. Scholkopf, A.J. Smola, Learning with Kernels, MIT Press, Cambridge, MA, 2002.
[8] J. Schurmann, Pattern Classification: A Unified View of Statistical and Neural Approaches, Wiley, New York, 1996.
[9] A.J. Smola, Regression estimation with support vector learning machines, Diplomarbeit, Technische Universitat Munchen, 1996.
[10] I. Steinwart, On the influence of the kernel on the consistency of support vector machines, J. Mach. Learn. Res. 2 (2001) 67–93.
[11] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer, New York, 1995.
[12] V.N. Vapnik, Statistical Learning Theory, Wiley, New York, 1998.
[13] V.N. Vapnik, Preface of Chinese edition of Statistical Learning Theory, Publishing House of Electronics Industry, Beijing, 2004.

Degang Chen received the M.S. degree from Northeast Normal University, Changchun, Jilin, China, in 1994, and the Ph.D. degree from Harbin Institute of Technology, Harbin, China, in 2000.

He has worked as a Postdoctoral Fellow with Xi'an Jiaotong University, Xi'an, China, from 2000 to 2002 and with Tsinghua University, Tsinghua, China, from 2002 to 2004. From 1994 to 2004, he had been a Teacher at Bohai University, Jinzhou, Liaoning, China, and since 2005, he has worked as a Teacher at North China Electric Power University. His research interests include fuzzy group, fuzzy analysis, rough sets and SVM.

Qiang He received the B.Sc. and M.Sc. degrees in mathematics from Hebei University, Baoding, China, in 2000 and 2003, respectively.

From 2003 till date, he worked as a Teaching Assistant at the Faculty of Mathematics and Computer Science, Hebei University. In 2004, he worked as a Research Assistant at the Department of Computing, Hong Kong Polytechnic University, Kowloon. His main research interests include inductive learning, genetic algorithms and statistical learning theory.

Xizhao Wang received the B.Sc. and M.Sc. degrees in mathematics from Hebei University, Baoding, China, in 1983 and 1992, respectively, and the Ph.D. degree in computer science from Harbin Institute of Technology, Harbin, China, in 1998.

From 1983 to 1998, he worked as a Lecturer, an Associate Professor, and a Full Professor at the Department of Mathematics, Hebei University. From 1998 to 2001, he worked as a Research Fellow at the Department of Computing, Hong Kong Polytechnic University, Kowloon. Since 2001, he has been the Dean and Professor of the Faculty of Mathematics and Computer Science, Hebei University. His main research interests include inductive learning with fuzzy representation, fuzzy measures and integrals, neuro-fuzzy systems and genetic algorithms, feature extraction, multi-classifier fusion, and applications of machine learning. So far, he has published over 30 international journal papers. He is an IEEE senior member and is an associate editor of IEEE Transactions on SMC part B.

Prof. Xizhao Wang is the General Co-Chair of the 2002, 2003, and 2004 International Conference on Machine Learning and Cybernetics, co-sponsored by IEEE SMC.