A Method to Construct the Mapping to the Feature Space for the Dot Product Kernels

Degang Chen[1], Qiang He[2], Chunru Dong [2], and Xizhao Wang[2]

[1] Department of Mathematics and Physics, North China Electric Power University, 102206, Beijing, P.R. China
chengdegang@263.net
[2] Department of Mathematics and Computer Science, Hebei University, Baoding, Hebei, 071002, P.R. China

Abstract. Dot product kernels are a class of important kernel in the theory of support vector machine. This paper develops a method to construct the mapping that map the original data set into the high dimensional feature space, on which the inner product is defined by a dot product kernel. Our method can also be applied to the Gaussian kernels. Via this mapping, the structure of features in the feature space is easy to be observed, and the linear separability of data sets in the feature space is studied. We obtain that any two finite sets of data with empty overlap in the original space will become linearly separable in an infinite dimensional feature space, and a sufficient and necessary condition is also developed for two infinite sets of data in the original data space being linearly separable in the feature space, this condition can be applied to examine the existences and uniqueness of the hyperplane which can separate all the possible inputs correctly.

1 Introduction

Support vector machine(SVM) is a new learning theory presented by Vapnik[1,2]. From the pattern recognition viewpoint, it can briefly be stated as follows. When a given sample set K is linearly separable, the separating hyperplane with the maximal margin, the optimal separating hyperplane, is constructed in the original space. When the sample set is linearly non-separating, the input vectors are mapped into the high-dimensional feature space through some kernel functions. Then in this high-dimensional feature space an optimal separating hyperplane is constructed. The inner product in the high-dimensional feature space is just the employed kernel, so the complex computing of inner product in the 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[3] and has been introduced as powerful tools for solving classification problems, at mean time the research on its theory and applications has drawn more and more attention in recent years.

However, if we only consider the computing of the inner product in the feature space, the kernel is enough, it is unnecessary to consider the mapping from the original data set to the feature space. But if we want to know more about the SVM, for example, analysis of the shape of mapped data in the feature space, consideration of the construction of features, and selection of optimal kernels with better generalization properties, the map-

D.S. Yeung et al. (Eds.): ICMLC 2005, LNAI 3930, pp. 918 – 929, 2006.

ping from the original data set to the feature space can not be ignored. In the existing statistical learning theory[6], there are mainly two approaches to obtain the mapping from the original data set to the feature space. One is to employed 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 the 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 they really exist. For the second approach, the 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 the original data set to the feature space with legible feature structure is clearly necessary from both of the theoretical and practical viewpoints.

As well known, dot product kernels are an important class of kernels in common use. The well known dot product kernels in the theory of SVM are homogeneous polynomial kernels, inhomogeneous polynomial kernels. Both the homogeneous polynomial kernels and inhomogeneous polynomial kernels map the original data set into a finite dimensional polynomial space(feature space) and the structures of features are clear(there is a whole field of pattern recognition research studying polynomial classifiers[4]). By using of the power series expansion of a dot product kernel, we can develop a mapping from the original dataset into a polynomial space(may not be finite dimensional) for every dot product kernel. Via this mapping, the structures of features are clear. This method can also be applied to the Gaussian kernels. Furthermore, the linear separability of data set is also investigated. 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. We also develop a sufficient and necessary condition for two infinite subclasses of the original data set being linear separable in the feature space, this condition offer a theoretical characterization to examine the existences and uniqueness of the hyperplane which can separate all the possible inputs correctly.

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. In section 4 we mainly discuss the separability of infinite sets in the feature space via our proposed mapping.

2 Kernels for SVM

In this paper we only consider the binary classification problem. Let $\{(x_1, y_1),...,(x_l, y_l)\} \subset R^n \times \{+1,-1\}$ 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 hyperplane $< w, x > +b = 0$ and $\delta > 0$ such that $< w, x > +b > \delta$ for

$x \in A$ and $< w, x > + b < -\delta$ for $x \in B$ (this definition is also suitable when A and B are infinite set), clearly $d(A,B) > 0$ holds when A and B are linear separable, and the separating hyperplane with the maximal margin, the optimal separating hyperplane, 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 hyperplane. Therefore, the classifier is given by $f(x) = sign(w^T \Phi(x) + b)$ and parameters w and b are obtained through the minimization of functional $\tau(w) = \frac{1}{2}\|w\|^2$ subject to $y_i(< w, x_i > + b) \geq 1$ for all $i = 1,...,l$. 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[6], which consists on expressing the inner product in Z as an evaluation of a kernel function in the input space $< \Phi(x), \Phi(y) >= k(x, y)$. 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 $f(x) = sign(\sum_{i=1}^{l} y_i \alpha_i k(x_i, x) + b)$, where coefficients $\{\alpha_i\}$ are obtained after a QP optimization of functional $L(w, b, \alpha) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{l} \alpha_i \{[< x_i, w > -b] y_i - 1\}$ which can be solved by the KKT complementarity conditions of optimization theory[3].

However, if we not only consider the computing of inner product in the feature space, but also aim to present deep insight to SVM such as analysis of the shape of mapped data in the feature space, consideration of the construction of features, and selection of optimal kernels with better generalization properties, we must deal with the mapping from original dataset into the feature space. As pointed in [6], there are mainly two approaches to develop the mapping. One is the utilization of the well known Mercer theorem. Suppose X is a nonempty 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 $(T_k f)(x) = \int_{\chi} k(x, x') f(x') d\mu(x')$ 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 $\Phi(x) = (\sqrt{\lambda_j}\psi_j(x))_{j=1,...,N_H}$. For this method, sometimes it is very difficult to compute the eigenvalues and eigenfunctions of T_k even 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 the structures of features are hardly to be observed.

Two kinds of kernels are always applied in SVM. 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 $k(x,x') = \exp(-\frac{\|x-x'\|^2}{2\sigma^2})$, other translation invariant kernels include B_n – splines kernels[7], Dirichlet kernels[6] and Periodic kernels[6]. A second important family of kernels can be efficiently described in term of dot product, i.e., $k(x,x') = k(<x,x'>)$. The well known dot product kernels are Homogeneous Polynomial Kernels $k(x,x') =< x,x'>^p$, inhomogeneous Polynomial Kernels $k(x,x') = (<x,x'>+c)^p$ with $c \geq 0$. Both Homogeneous Polynomial Kernels and inhomogeneous Polynomial Kernels map the input set into a finite dimensional Polynomial space. In [11] 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 $k_c(x,x') = \frac{1-<x,x'>^p}{(1-<x,x'>)^p}$, for every $x, x' \in U_n$, $p \in N - \{1\}$, via an infinite Polynomial kernel, the input dataset 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 dataset 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 the original space will become linearly separable in the feature space.

For the dot product kernels, the following theorem is always useful.

Theorem 1.[8] A function $k(x,x') = k(<x,x'>)$ defined on an infinite dimensional Hilbert space, with a power series expansion $k(t) = \sum_{n=0}^{\infty} a_n t^n$ is a positive definite kernel if and only if for all n, we have $a_n \geq 0$.

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

Suppose $k(x,x') = k(<x,x'>)$ is a dot product kernel on $X \subset R^n$ with the power series expansion $k(<x,x'>) = \sum_{n=0}^{\infty} a_n <x,x'>^n$. For every $x \in X$, define C_n to map $x \in X$ to the vector $C_n(x)$ whose entries are all possible n th 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 $<C_n(x),C_n(x')>=<\Phi_n(x),\Phi_n(x')>=<x,x'>^n$. This fact can be found in [6] and is well known for the Homogeneous Polynomial Kernels $k(x,x')=<x,x'>^p$.

Define $\Phi(x)=(a_0,\sqrt{a_1}\Phi_1(x),...,\sqrt{a_n}\Phi_n(x),...)$, then we have $<\Phi(x),\Phi(x')>=k(x,x')$. 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 our proposed Φ. The entries of $\Phi(x)$ is constructed by the entries of x, thus the structure of the 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 $\Phi(X)$.

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 $k(x,x')=\sum_{n=0}^{n_0}a_n<x,x'>^n$, thus $k(x,x')$ is just the weighted sum of some Homogeneous Polynomial Kernels, and the feature space is a finite dimensional Homogeneous Polynomial Kernels. However, for $k(x,x')=<x,x'>^n$, 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 2. 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 hyperplane in the feature space relative to the kernel $k(x,x')=<x,x'>^2$ can not distinguish x and y. Similar cases will appear frequently when n is an even. If we select a weighted sum form kernel predigest satisfying $a_1\neq 0$, then the entries of x is a part of $\Phi(x)$, thus we can avoid this case.

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

Theorem 3. Suppose $\{x_1,...,x_m\}\subset X$ satisfying $x_i\neq 0$, $x_i\neq x_j$ if $i\neq j$, then there is a dot product kernel $k(x,x')=\sum_{n=0}^{n_0}a_n<x,x'>^n$ such that $\Phi(x_1),...,\Phi(x_n)$ are linear independent.

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

Let $f_i(x) = a_{i1} + a_{i2}x + ... + a_{in}x^{n-1}$, $i = 1,...,m$. If $i \neq j$, then $x_i \neq x_j$, we have $f_i(x)$ 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 roots. Thus there exists $n_0 \in N$ such that any two of $\{f_i(n_0) : i = 1,...,m\}$ are different. Let $\beta_i = \{1, f_i(n_0),..., f_i^{m-1}(n_0)\}$, $i = 1,...,m$, then we have $\beta_1, \beta_2,..., \beta_m$ are linear independent.

Suppose $\alpha_1\Phi(x_1) + \alpha_2\Phi(x_2) + ... + \alpha_m\Phi(x_m) = 0$, then $\sum_{i=1}^{m} \alpha_i a_{i1}^{l_1} a_{i2}^{l_2} ... a_{in}^{l_n} = 0$, $l_1 + l_2 + ... + l_n \leq m-1$, $l_1, l_2,..., l_n \in N \bigcup \{0\}$, we have $\sum_{i=1}^{m} \alpha_i f_i^n(n_0) = 0$, this implies $\alpha_1\beta_1 + ... + \alpha_m\beta_m = 0$, thus every $\alpha_i = 0$ and $\Phi(x_1),...,\Phi(x_n)$ are linear independent.

In the proof of Theorem 3 we choice the kernel as $k(x,x') = \dfrac{1 - < x,x' >^m}{1 - < x,x' >}$ in order to predigest the proof. However, every kernel $k(x,x') = \sum_{n=0}^{n_0} a_n < x,x' >^n$ satisfying $n_0 \geq m-1$ and $a_n > 0$ for $n \leq m-1$ satisfies the condition in Theorem 3.

Suppose Φ is a mapping relative to a kernel $k(x,x')$ such that $\Phi(x_1),...,\Phi(x_n)$ are linear independent, A and B are two nonempty subsets of X and $A \bigcap B = \phi$, then we have $\Phi(X) = \Phi(A) \bigcup \Phi(B)$ and $\Phi(A) \bigcap \Phi(B) = \phi$. $\Phi(x_1),...,\Phi(x_n)$ 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 the two convex hulls of A and B have empty overlap, notice these two convex hulls are compact, so $\{\Phi(x_1),...,\Phi(x_l)\}$ and $\{\Phi(x_{l+1}),...,\Phi(x_m)\}$ are linear separable in the feature space. Thus we can derive the following fact.

Theorem 4. Suppose $\{(x_1, y_1),...,(x_l, y_l)\} \subset X \times \{+1\}$, $\{(x_{l+1}, y_{l+1}),...,(x_m, y_m)\} \subset X \times \{-1\}$, 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 the feature space.

Suppose $k(< x,x' >) = \sum_{n=0}^{\infty} a_n < x,x' >^n$ 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 $k(x,x') = (1 - (< x,x' >))^{-1}$ [6]and our proposed infinite polynomial kernel $k_c(x,x') = \dfrac{1 - < x,x' >^p}{(1 - < x,x' >)^p}$ [11]. The following theorem implies the feature space relative to such kernels is infinite dimensional.

Theorem 5. Suppose $\{x_1,...,x_m\} \subset X$ satisfies $x_i \neq 0$ for $i = 1,2,...,m$, $x_i \neq x_j$ if $i \neq j$, Φ is the mapping relative to $k(< x, x' >) = \sum_{n=0}^{\infty} a_n < x, x' >^n$ such that every $a_n > 0$, then $\Phi(x_1),...,\Phi(x_n)$ are linear independent.

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

Let $f_i(x) = a_{i1} + a_{i2}x + ... + a_{in}x^{n-1}$, $i = 1,...,m$. Then there exists $n_0 \in N$ such that any two of $\{f_i(n_0) : i = 1,...,m\}$ are different. Let $\beta_i = \{1, f_i(n_0),..., f_i^{m-1}(n_0)\}$, $i = 1,...,m$, then we have $\beta_1, \beta_2,.., \beta_m$ are linear independent. But by $\sum_{i=1}^{m} \alpha_i a_{i1}^{l_1} a_{i2}^{l_2} ... a_{in}^{l_n} = 0$ we have $\alpha_1\beta_1 + ... + \alpha_m\beta_m = 0$, this is a contradiction. Thus we have $\Phi(x_1),...,\Phi(x_n)$ are linear independent.

For $\{x_1,...,x_m\} \subset X$, Theorem 3 implies there exists a finite dimensional feature space such that the images of $\{x_1,...,x_m\}$ are linear independent in this feature space, while Theorem 5 implies the images of $\{x_1,...,x_m\}$ are linear independent in the feature space relative to a kernel satisfying every coefficient in its power series is positive, so these two theorems are different. For the kernel satisfying every coefficient in its power series is positive, similar to Theorem 4 we have the following result.

Theorem 6. Suppose $\{(x_1, y_1),...,(x_l, y_l)\} \subset X \times \{+1\}$, $\{(x_{l+1}, y_{l+1}),...,(x_m, y_m)\} \subset X \times \{-1\}$, then they are linear separable in every feature space relative to a kernel satisfying 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 the selection of feature space does not influence the linear independence of a finite class of data in the feature space.

Theorem 7. Suppose $\{x_1,...,x_m\} \subset X$ satisfies $x_i \neq 0$ for $i = 1,2,...,m$, $x_i \neq x_j$ 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 every coefficient in its power series is positive.

Proof. If $M = (k(x_i, x_j)) = (< \Phi(x_i), \Phi(x_j) >)$ has not full rank, then there exists $\alpha_1, \alpha_2,..., \alpha_m$ satisfying at least one of them is not equal to zero such that $\sum_{l=1}^{m} \alpha_l < \Phi(x_l), \Phi(x_i) >= 0$, $i = 1,..., m$. So we have

$< \alpha_i \Phi(x_i), \sum_{l=1}^{m} \alpha_l \Phi(x_l) >= 0$, $i = 1,...,m$ which implies $< \sum_{i=1}^{m} \alpha_i \Phi(x_i), \sum_{l=1}^{m} \alpha_l \Phi(x_l) >= 0$, thus $\sum_{i=1}^{m} \alpha_i \Phi(x_i) = 0$ and $\Phi(x_1),...,\Phi(x_n)$ are linear dependent. Hence $M = (k < x_i, x_j >) =< \Phi(x_i), \Phi(x_j) >$ has full rank.

If Φ' is another mapping that project X into a different feature space, then it is easy to prove $\Phi'(x_1), \Phi'(x_2),..., \Phi'(x_m)$ 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 8. If Φ_2 is the extension of Φ_1, then $\Phi_1(x_1),...,\Phi_1(x_n)$ are linear independent implies $\Phi_2(x_1),...,\Phi_2(x_n)$ 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 an unit vector, i.e., $\|x\| = 1$, then $\|x - x'\|^2 =< x - x', x - x' >= 2 - 2 < x, x' >$, thus the Gaussian kernels $k(x, x') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2})$ have an equivalence expression as dot product kernels as $k(x, x') = \exp(\frac{< x, x' > -1}{\sigma^2})$, and we can construct the mapping for the Gaussian kernels by its power series by our proposed method. In [6] 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 $\Phi_G(x_1),...,\Phi_G(x_m)$ are linear dependent for $\{x_1,...,x_m\} \subset X$, this statement is very important for analysis of the properties of Gaussian kernels . By Theorem 5 we can also get this conclusion and we propose a new straight proof for this result, our proof is different with the original one in [13].

For a finite data set $\{x_1,...,x_m\} \subset X$, $\Phi(x_1),...,\Phi(x_n)$ are linear independent implies any binary partition of $\{x_1,...,x_m\}$ are linear separable in the feature space. So $\Phi(x_1),...,\Phi(x_n)$ being linear independent is a sufficient condition of $\{x_1,...,x_m\}$ being linear separable in the 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 $\{x_1,...,x_m\}$ in stead of separating all its possible binary partitions. The equivalence description of linear separability of a binary classifications problem by using of kernel is a meaningful problem.

4 On the Linear Separability of Infinite Data Sets in Feature Space

In this section we mainly discuss the linear separability of infinite data sets in feature space. At first glance, it is unnecessary to consider infinite data sets since the 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 the theoretical viewpoint, the following three arguments indicate it is meaningful to investigate the linear separability of infinite data sets.

First, separating two finite sets linearly is equivalence 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 data may be infinite even the samples are infinite, for instance, if we take stature as a feature with value range 0.5 to 2.5 meter, then every number between 0.5 and 2.5 is possible to be the stature of somebody. So after we construct a learning machine based on finite independent and identically distributed samples, the possible data we deal with by this machine is always drown from an infinite set and we can not exactly forecast its detail structure, i.e., exact values of the possible data taking for every feature, this also need to take account of all possible cases drown according to a probability distribution. At last, for a practical binary problem, certainly we desire to know the existence and uniqueness of optimal hyperplane that can separate all the possible data without misclassification, this also inspires us to consider all the possible data.

Thus it is necessary to investigate the linear separability of infinite data sets at least from the theoretical viewpoint, and such investigation can offer guidance to improve algorithm for practical problems.

For any two finite data sets, by our discussion in Section 3 there must exists a feature space relative to a dot product kernel such that they are linear separable in feature space, and the optimal hyperplane in the feature space is always available. For the infinite data set this statement may not hold, we have the following sufficient and necessary condition to characterize the linear separability of two infinite data sets.

Theorem 9. Suppose $X \subset R^n$ is compact and $X = A \bigcup B$, $A \bigcap B = \phi$. Then there exists a feature space relative to a dot product such that $\Phi(A)$ and $\Phi(B)$ are linear separable in feature space if and only if the crowed point sets of A and B have empty overlap, i.e., the boundary points set of A and B is empty.

Proof. Without losing universality, we assume X is a subset of the open unit ball. We select Vovk's infinite polynomial kernel $k(x, x') = (1 - (< x, x' >))^{-1}$ in the following proof, Φ is the mapping relative to $k(x, x') = (1 - (< x, x' >))^{-1}$.

$\Rightarrow$ Since X is compact, we know the crowed points of X are still in X, thus the crowed points of $\Phi(X)$ are still in $\Phi(X)$ by $\Phi(X)$ is compact. If the 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 $\Phi(B)$ can not 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 $\Phi(B)$ are empty, thus they are linear separable in the feature space and $\Phi(A)$ and $\Phi(B)$ are linear separable in the feature space.

Other kernels $k(<x,x'>)=\sum_{n=0}^{\infty} a_n <x,x'>^n$ satisfies for every $n_0 \in N$ there exists $n > n_0$ such that $a_n > 0$ can also be employed to prove this theorem.

For a binary pattern recognition problem, if there is a hyperplane which can not only separate the training simple but also can classify every possible 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 9 develops a sufficient and necessary condition under which a binary pattern recognition problem is possible to be solved totally, i.e., for every sample of one class, there exists a sufficient small neighborhood of this sample satisfying none sample of another class is in this neighborhood. Thus we can conclude that for a binary pattern recognition problem, if it can be solved totally, then generally the selection of optimal separating hyperplane is not unique, if it can not be solved totally, then the optimal separating hyperplane does not exist. The following figure illustrates our idea of Theorem 5.

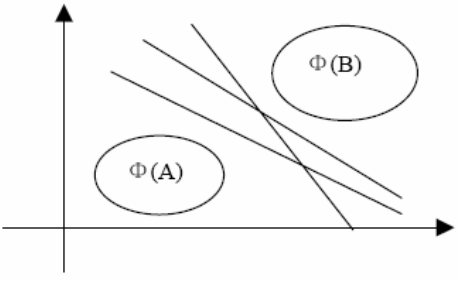

Fig. 1. If X is compact and $X = A \bigcup B$, $A \bigcap B = \phi$, all the possible data are in X. If the crowed point sets of A and B have empty overlap, then $\Phi(A)$ and $\Phi(B)$ are linear separable in the feature space as shown in above figure. Since every separating hyperplane can classify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

As pointed out in [12], since one has to make assumptions about the structure of the data(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 output[6]. Most classical classification algorithms rely, implicitly or explicitly, on such an assumption(e.g. nearest-neighbors classifiers, and the

simplest possible justification for large margins in SVM in [6]). Applying this assumption to the binary pattern recognition problems, it just implies the crowed points of the two classes have an empty overlap, thus the optimal separating hyperplane in the feature space always exists and is not unique.

If the binary pattern recognition problems do not satisfy this assumption, i.e., the two classes have conjunct crowed points, then the optimal separating hyperplane that can separate all the data without misclassification is not available. By 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 hyperplane should pass through the crowed points. We employ the following simple example to illustrate our idea.

Example 3. 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 can explain this example straightforward.

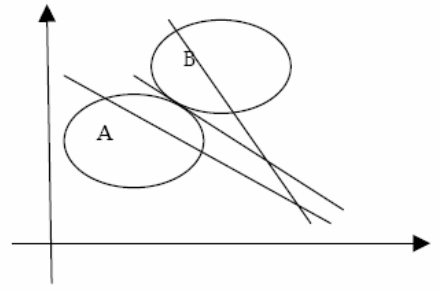

Fig. 2. To separate two tangent ellipses by a line, clearly the tangent is the best selection

Clearly the conjunct points may not be unique, and the number of conjunct points will influence the selection of separating hyperplane. We omit detail discussion on this topic here and will focus on it in detail in another paper.

Acknowledgements

This paper is supported by a Foundation of North China Electric Power University, the National Natural Science Foundation of china(NSFC60473045) and the Natural Science Foundation of Hebei Province (603137)

References

1. Vapnik V. N.: The Nature of Statistical Learning Theory. New York: Springer- Verlag(1995)
2. Vapnik V. N.: Statistical Learning Theory. New York: Wiley(1998)

3. Burges C.: A Tutorial on Support Vector Machines for Pattern Recognition, Data Mining and Knowledge Discovery, 2(2)(1998)121-167
4. Schurmann J.: Pattern Classification: A Unified View of Statistical and Neural Approaches. Wiley, New York(1996)
5. Micchelli C. A.: Algebraic Aspects of Interpolation, Proceedings of Symposia in Applied Mathematics, 36(1986) 81-102
6. Scholkopf B., Smola A. J.: Learning with Kernels, MIT Press, Cambridge, MA(2002)
7. Smola A. J.: Regression Estimation with Support Vector Learning Machines, Diplomarbeit, Technische Universitat Munchen(1996)
8. Schoenberg I. J.: Positive Definite Functions on Spheres, Duke Mathematical Journal, 9(1942) 96-108
9. Steinwart I.: On the Influence of the Kernel on the Consistency of Support Vector Machines, Journal of Machine Learning Research 2(2001)67-93
10. Saunders C., Stitson M. O., Weston J., Bottou L., Scholkopf B., and Smola A. J.: Support Vector Machine Reference Manual. Technical Report CSD-TR-98-03, Department of Computer Science, Royal Holloway, University of London, Egham, UK(1998)
11. Chen Degang, He Qiang, Wang Xizhao.: The infinite polynomial kernel for support vector machine, Lecture Notes in Artificial Intelligence 3584(2005): 267-275
12. Matthias Hein, Olivier Bousquet, Bernhard Scholkopf.: Maximal margin classification for metric spaces, Journal of Computer and System Sciences 71(2005)333-359
13. C. A. Micchelli.: Algebraic aspects of interpolation. Proceedings of Symposia in Applied Mathematics 36(1986)81-102