The Infinite Polynomial Kernel for Support Vector Machine

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

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

Abstract. This paper develops an infinite polynomial kernel k_c for support vector machines. We also propose a mapping from an original data space into the high dimensional feature space on which the inner product is defined by the infinite polynomial kernel k_c. Via this mapping, any two finite sets of data in the original space will become linearly separable in the feature space. Numerical experiments indicate that the proposed infinite polynomial kernel possesses some properties and performance better than the existing finite polynomial kernels.

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 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 tool for solving classification problems. In the mean time the research on SVM theory and applications has drawn more and more attention in recent years. As well known that kernel is one of the most important concepts in the theory of SVM and many efforts have been concentrated to the research of kernels. The well known kernels in the theory of SVM are homogeneous polynomial kernels, inhomogeneous polynomial kernels, Gaussian radial basis function kernels, sigmoid kernels and $B_n -$spline 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

X. Li, S. Wang, and Z.Y. Dong (Eds.): ADMA 2005, LNAI 3584, pp. 267 – 275, 2005.

studying polynomial classifiers [4]), but it is possible that for a fixed polynomial kernel there exists a data set which is not separable in the feature space relative to this kernel since the feature space is finite dimensional. In the mean time the Gaussian radial basis function kernels map the original data set into an infinite dimensional space and any finite data set is linear separable in the feature space with respect to this kernel [5], but the structures of the features relative to the Gaussian radial basis function kernels are difficult to analysis. This statement suggests us to consider infinite polynomial kernels for SVM. In this paper we propose an infinite polynomial kernel on the open unit ball and study the map with respect to this kernel which map the original data set into the feature space, we also prove that by this map the images of any finite data set are linear independent in the feature space, this implies any two finite subclasses of the original data set are linear separable in the feature space. Our experiment indicates that this infinite polynomial kernel can really reduce the number of support vectors thus it possesses better properties than the finite polynomial kernel. Thus this kernel can be applied to solve practical problems.

The rest of this paper is organized as follows. A brief review of the theory of SVM will be described in Section 2. The infinite Polynomial Kernels in the open unit ball will be derived in Section 3. Experiments are presented in Section 4. Some concluding remarks are given in Section 5.

2 Kernels for SVM

Let $\{(x_1, y_1),...,(x_l, y_l)\} \subset R^n \times \{+1,-1\}$ be a training set. 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 $< \Phi(x), \Phi(y) >= k(x, y)$. 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].

From the above analysis it is clear that the kernel play a key role in the application of SVM, thus a deep insight to the structure of kernels is both of theoretical and practical important. There are two approaches to characterize the kernel [6]. First it can be believed as inner product in a Reproducing Kernel Hilbert Space [6]. On the other hand it is a symmetric real-valued function satisfying the well known Mercer Theorem [6]. The latter statement is always employed to examine a function to be a kernel.

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 are 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. This implies it is possible that two classes of inputs may be non-separable in the feature space for a fixed Polynomial Kernel. 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 that many kinds of dot product kernels can be considered in SVM.

3 The Infinite Polynomial Kernels in the Open Unit Ball

Since both Homogeneous Polynomial Kernels and inhomogeneous Polynomial Kernels map the input set into a finite dimensional Polynomial space and they cannot linearly separate all the data set in the feature space, they are not very satisfied at least from the theoretical viewpoint even they perform well in some practical problems. In this paper, to overcome the above weakness, we consider a class of infinite Polynomial Kernels in the open unit ball $U_n = \{x \in R^n : \|x\| < 1\}$ which can make any finite data set in U_n linear separable in the high dimensional feature space.

Theorem 2. For every $x, x' \in U_n$, $p \in N - \{1\}$, define $k_c(x,x') = \frac{1-<x,x'>^p}{(1-<x,x'>)^p}$, then k_c is a kernel.

Proof. By $x, x' \in U_n$ we have $|<x,x'>| < 1$. Let $k_c(t) = \frac{1-t^p}{(1-t)^p}$, then we have $k_c(t) = (1+t+...+t^{p-1})(\sum_{k=0}^{\infty} t^k)^p$ for $|t| < 1$, by Theorem 1 we know $k_c(<x,x'>)$ is a kernel.

Suppose $k_c(t)=\sum_{k=0}^{\infty}a_k t^k$, then $a_k=\frac{k_c^{(k)}(0)}{k!}$. For every $x \in U_n$, define C_k to map $x \in U_n$ to the vector $C_k(x)$ whose entries are all possible k th degree ordered products of the entries of x , and define Φ_k by compensating for the multiple occurrence of certain monomials in C_k by scaling the respective entries of Φ_k with the square roots of their numbers of occurrence. Then, by the construction of C_k and Φ_k , we have $<C_k(x),C_k(x')>=<\Phi_k(x),\Phi_k(x')>=<x,x'>^k$.

Define $\Phi(x)=(1,\sqrt{a_1}\Phi_1(x),...,\sqrt{a_k}\Phi_k(x),...,)$, then we have $<\Phi(x),\Phi(x')>=k_c(x,x')$. The feature space with respect to $k_c(x,x')$ can be selected as the Hilbert space generated by $\Phi(U_n)$. The following theorem implies this space is infinite dimensional.

Theorem 3. Suppose $\{x_1,...,x_m\}\subset U_n$ satisfying $x_i \neq x_j$ if $i \neq j$, 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 $l_1,l_2,...,l_n \in N \cup \{0\}$.

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.

Furthermore by Theorem 3 we have the following conclusions.

Theorem 4. Suppose $\{(x_1,y_1),...,(x_l,y_l)\}\subset U_n \times \{+1\}$, $\{(x_{l+1},y_{l+1}),...,(x_m,y_m)\}\subset U_n \times \{-1\}$, then $\Phi(x_1),...,\Phi(x_l)$ and $\Phi(x_{l+1}),...,\Phi(x_m)$ are linear separable in the feature space.

Proof. $\Phi(x_1),...,\Phi(x_n)$ 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 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 for any finite data set the optimal hyperplane in the feature space is always available.

Theorem 5. Suppose $\{x_1,...,x_m\} \subset U_n$ satisfying $x_i \neq x_j$ if $i \neq j$, then the Gram matrix $M = (k_c < x_i, x_j >) =< \Phi(x_i), \Phi(x_j) >$ has full rank.

Proof. If $M = (k_c < x_i, x_j >) =< \Phi(x_i), \Phi(x_j) >$ 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 $M = (k_c < x_i, x_j >) =< \Phi(x_i), \Phi(x_j) >$ has full rank.

The feature space with respect to $k_c(< x, x' >)$ is not uniqueness, and Theorem 5 indicates that the selection of feature space(mapping) does not influence the linear independence of a finite class of data in the feature space. By the proof of Theorem 3 we can easily get the following conclusion for the finite Polynomial kernels.

Theorem 6. Suppose $\{(x_1, y_1),...,(x_l, y_l)\} \subset U_n \times \{+1\}$, $\{(x_{l+1}, y_{l+1}),...,(x_m, y_m)\} \subset U_n \times \{-1\}$, then there exists $p \in N$ such that their images are linear separable in the feature space with respect to the kernel $< x, x' >^p$ or $(< x, x' > +1)^p$.

The feature space with respect to every finite Polynomial kernel can be embedded into the feature space with respect to the kernel $k_c(x, x')$ as a subspace, this means there has more different features in the feature space with respect to the kernel $k_c(x, x')$ to be applied to pattern recognition and all these features are constructed by the entries of the input vector. Thus the kernel $k_c(x, x')$ possesses the advantages of Gaussian radial basis function kernels and Polynomial kernels, i.e., it can linearly separate any finite data set and constructions of features are clear, we hope it may perform well in practical problems than the finite Polynomial kernels, we will examine this statement by the experiments in the following section.

4 Experiments

In this section, for the purpose of examining infinite polynomial kernel, we would like to select four databases from machine learning repository (UCI). For these databases, the performance based on new kernel in previous section and finite polynomial kernel will be summarized and compared. Optdigits database includes 5620 cases with 10 classes, 1119 cases are randomly selected to demonstrate. Since the SVM is only for two-class classification problems in this paper, we unite the cases to one class, which belong to class (0,2,4,6,8), and the remaining cases are used as the other class. The four databases' characters are shown in table 1. Applying SVM Toolbox (http://www.isis.ecs.soton.ac.uk/isystems/kernel/svm.zip) to the original data of the four selected databases, one can obtain the optimal separating hyper-planes. The results of these experiments are given in table 2 to 13, where 80% of the databases are randomly selected as the training sets and the remaining 20% as the testing sets. For different types of kernels, the tables show the parameters and the corresponding performance. It is worth noting that the experimental results also depend on the many parameters chosen in the SVM Toolbox.

From tables 2,4,6,8, one can see that the training and testing accuracy are indeed enhanced using infinite polynomial kernel. However, the improvement is not significant. We speculate that the reason is that (1) the data is not enough and (2) database is linear separable very much.

Table 1. The characters of databases

Database Name	Number of samples	Number of features	Category of features
rice	105	5	Numerical
sonar	208	60	Numerical
pima	768	8	Numerical
optdigits	1119	64	Numerical

Table 2. Experiment results for rice database

	infinite polynomial kernel			finite polynomial kernel		
P	Training Accuracy	Testing Accuracy	SV Number	Training Accuracy	Testing Accuracy	SV Number
2	100	93.75	70	100	90.625	73
4	100	93.75	69	100	93.75	73
8	100	96.875	58	100	93.75	72
16	100	96.875	32	100	96.875	40
32	100	96.875	18	100	96.875	17
64	100	96.875	8	100	96.875	9

Table 3. Percentage of common support vector for various kernels for rice database

P	2	4	8	16	32	64
Infinite polynomial kernel	100	100	100	100	94.4	100
Finite polynomial kernel	95.9	94.5	80.6	80.0	100	88.9

From tables 3,5,7,9, one important feature was observed: two types of kernels use approximately the same set of support vectors, but the number of support vectors for infinite polynomial kernel is small in a way(only two cases happen that the number of support vectors for infinite polynomial kernel is bigger than the number of support vectors for finite polynomial kernel), this implies the number of support vectors is really reduced by the infinite polynomial kernel. Noticed that for the support vectors machines, less support vector means better performance of the SVM, so SVM with infinite polynomial kernel developed in this paper have better properties than SV machines with finite polynomial kernel.

Table 4. Experiment results for sonar database

	infinite polynomial kernel			finite polynomial kernel		
P	Training Accuracy	Testing Accuracy	SV Number	Training Accuracy	Testing Accuracy	SV Number
2	100	78.571	159	100	78.571	161
4	100	80.952	124	100	78.571	129
8	100	80.952	91	100	78.571	93
16	100	83.333	69	100	80.952	69
32	100	78.571	65	100	78.571	67
64	100	85.714	67	100	85.714	67

Table 5. Percentage of common support vector for various kernels for sonar database

P	2	4	8	16	32	64
Infinite polynomial kernel	99.4	100	100	100	100	100
Finite polynomial kernel	98.1	96.1	97.8	100	97	100

Table 6. Experiment results with infinite polynomial kernel for pimar database

	infinite polynomial kernel			finite polynomial kernel		
P	Training Accuracy	Testing Accuracy	SV Number	Training Accuracy	Testing Accuracy	SV Number
2	76.384	80.519	562	74.675	78.631	614
4	76.221	80.519	561	74.675	79.268	614
8	76.71	80.519	562	75.974	80.126	614
16	77.036	80.519	560	77.036	80.519	560
32	77.036	80.519	560	76.873	80.519	560
64	77.036	80.519	557	76.873	80.519	614

Table 7. Percentage of common support vector for various kernels for pima database

P	2	4	8	16	32	64
Infinite polynomial kernel	100	100	100	100	100	100
Finite polynomial kernel	91.5	91.4	91.5	100	100	90.7

Table 8. Experiment results with for optdigit database

	infinite polynomial kernel			finite polynomial kernel		
P	Training Accuracy	Testing Accuracy	SV Number	Training Accuracy	Testing Accuracy	SV Number
2	96.745	94.737	262	95.398	92.982	891
4	99.327	96.491	219	98.653	94.982	234
8	100	96.053	891	99.888	96.053	891
16	100	96.491	139	100	96.491	142
32	100	96.053	891	100	96.053	116
64	100	96.053	891	100	96.053	891

Table 9. Percentage of common support vector for various kernels for optdigit database

P	2	4	8	16	32	64
Infinite polynomial kernel	100	97.7	100	100	100	100
Finite polynomial kernel	29.4	91.5	100	97.9	13	100

5 Conclusion

The purpose of this paper is to present infinite polynomial kernel for SVM. By our theoretical analysis this kernel possesses better properties than the existing finite polynomial kernel. Our experiments results almost support our opinion. The infinite polynomial kernel can be applied to practical problems. Further research to the properties and applications of infinite polynomial kernel will be our future work.

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, vol. 2, no. 2, 1998

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: 81-102, 1986
6. Scholkopf, B. and 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: 96-108, 1942