第24卷 第4期 广西师范大学学报:自然科学版 Vol. 24 No. 4
2006年12月 Journal of Guangxi Normal University:Natural Science Edition Dec. 2006

Framework for Choice of Kernels

WAN Hai-ping, HE Hua-can

(School of Information Engineering, Beijing University of Post and Telecommunication, Beijing 100876, China)

Abstract: How to select a proper kernel is a challenging problem. From the view of manifold learning and conformal map, an exploratory way on how to select a kernel for a particular machine learning task is proposed. Then an experiment is performed to verify our scheme and some initial inclusions are drawn.

Key words: Kernel method; manifold learning; conformal map

CLC number: TP18 **Document code**: A **Article ID**: 1001-6600(2006)04-0050-04

0 Introduction

In many machine learning tasks such as pattern classification or clustering, one is often confronted with high dimensionality data. While the high dimensionality may provide more information for the learning task, it also incurs the curse of dimensionality at the same time: the higher the dimensionality is, the more data samples one needs in order to explore its distribution tendency. The latter quantity is often an exponential expression of the former one. One will be dejected to find it is computationally infeasible at all in the end with the increase of dimensionality. So dimensionality reduction technique is often performed before the learning task begins. However, the representation ability of low dimension space is weaker than that of the higher dimension space; this is really a dilemmatic situation usually afflicting people concerned.

Kernel methods provide a small trick which alleviates the dilemma to some extent. While it maps the data to a higher dimensionality space, the space is used implicitly via a kernel function defined on the original space to elude the curse of dimensionality. This is a useful trick, for many canonical algorithms, for example, multidimensional scaling (MDS), Fisher discriminant analysis and principal component analysis can all be generalized to their nonlinear forms by substituting a kernel function for the inner product[1].

In the last few years, kernel methods performed very well in many machine learning tasks such as text categorization, time-series prediction, and protein homology detection. There are several popular kernel functions in practice such as polynomial kernels, Gaussian kernels, and sigmoid kernels. Also new kernels can be generated from existing ones. But how to judge if a specific kernel function is proper for a particular learning task? It remains a challenging problem in the study of kernel methods. Usually it depends on domain knowledge or experience of an expert. There is often no directive or automatic method to seek a proper kernel function. We know that a kernel function is defined as an inner product in the feature space:

$$K(x,z)=\langle\varphi(x)\cdot\varphi(z)\rangle. \tag{1}$$

Where φ is a map from X to an inner product feature space F, $x,z\in X$, Although the mapping φ and the

Received date: 2006-05-31

Foundation item: Foundation of Ministry of Education (MZ115-022)

Biography: WAN Hai-ping (1977—), male, born in Jinan, Shandong, Postdoctor of Beijing University of Post and Telecommunication.

space F is often not explicitly defined, one cannot turn aside in order to evaluate the quality of the kernel it deduces. We discuss this issue in the next section.

1 Conformal Map View of Kernels

Many data projection techniques aim to preserve the global structure. We argue that it is unnecessary for many practical applications. Consider the spherical surface S^2 described by $x^2+y^2+z^2=1$, it is easy to see that in either the upper or the lower half spherical surface z can be uniquely determined by two coordinates x, y. But there is not one binary continuous function that can uniquely determine a point on the whole spherical surface. Think about another scenario. What if one sees only a half of an image of a lion? While this mapping φ loses about a half of the whole information, in most cases he is still able to recognize that it is a lion although there is only some local fitting preserved. But if the whole image is distorted seriously by some mapping φ, it will muddle him. In many perception scenarios, we can still figure out the whole image even there is only some locality fitting preserved. So a good mapping should not lose meaningful local structure of the data. The above examples show that sometimes it is unnecessary, expensive or impossible to construct a uniform coordinate system for all the data samples we are examining, and some local fitting will be enough for the learning task.

Now it is widely accepted that although the sampled data is high dimensional at first glance, it is most probably that it lies on a low dimensional manifold M[2]. From the view of manifold theory, we can use different coordinates for different parts of the data samples. For example, in a local part the samples may present a structure that is easy to capture than capturing a global structure for all the samples. A good mapping should not contort the embedded manifold too much to lose useful information, especially the dimensionality of M. If some manifold learning algorithm can be conducted in the feature space F and the dimensionality of the manifold M can be estimated, then the effects of the manifold learning algorithm applied to different kernel functions can be compared to seek the majorities of the results, thereby we get some directive knowledge on how to select a kernel function.

2 Manifold Learning in the Kernel Space

There have been several manifold learning algorithms proposed. Isomap[3], locally linear embedding[4], graph Laplacian eigenmap[5] all utilize local neighborhood information to build a global embedding of the manifold. Isomap is a generalization of the classical MDS. MDS seeks a low-dimensional embedding based on pairwise similarity between data points, where Euclidean distance is often used as a measure of similarity. The basic idea in Isomap is to use geodesic distance on a neighborhood graph in the framework of MDS. It first constructs a symmetric adjacency graph using criteria such as nearest neighborhoods or ε-ball neighborhoods. Weights of the edges are signed as Euclidean distances between neighboring pointes. Then geodesic distance between two different points is approximated by the length of the shortest path connecting them in the graph, which can be found by several existing shortest path searching algorithms. Finally, MDS is applied to the distance matrix to find the low-dimensional embedding.

Now we consider how to conduct Isomap in the kernel feature space F. Let $X=\{X_1,\cdots,X_n\}$ be the data samples, $\varphi(X)$ is its image in the kernel space F.

First, Construct an adjacency graph G over all data points in $\varphi(X)$ by connecting points $\varphi(x_i)$ and $\varphi(x_j)$ if the distance $d(i,j)$ is closer than ε. Assign $d(i,j)$ to the corresponding edge length, where $d^2(i,$

$j)=K(x_i,x_i)-2K(x_i,x_j)+K(x_j,x_j)$, Let $D=\{d(i,j)\}$.

Secondly, approximate the geodesic distance by the length of the shortest path in graph G. We use the Bell-Man ford algorithm[6] to get the matrix $\boldsymbol{D}'$ which contains the pairwise geodesic distance between all points in G.

Then seek a d-dimensional embedding space Y to minimize the cost function

$$E=\|\tau(\boldsymbol{D}')-\tau(\boldsymbol{D}_Y)\|_{L^2} \tag{2}$$

Where L^2 is the Frobenius matrix norm and τ is the operator converting distances to inner products, $\boldsymbol{D}_Y$ is the matrix of Euclidean distances $\{d_Y(ij)\}$. The global minimum is achieved as follows: Suppose λ_p is the p-th largest eigenvalue of the matrix $\tau(\boldsymbol{D}')$, and v_{p_i} is the i-th component of the corresponding eigenvector. Set the p-th component of y_i equal to $\sqrt{\lambda_p}v_p^i$[7].

The dimensionality of the data d can be estimated from the decrease in errors as d increases. Suppose that we have several alternative kernel functions in a particular learning task, then for each kernel function we perform the above manifold learning algorithm in its corresponding kernel feature space. Then we seek the majority of the experiments, regard that it is the true dimensionality of the embedding manifold, and we get some directive information on how to select a proper kernel.

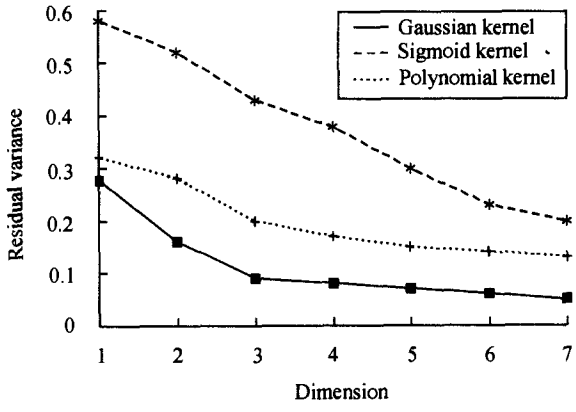

Fig. 1 Relationship between dimensionality and residual variance of 3 kernels

Their residual variances are listed in figure 1.

We conduct an experiment to verify our scheme. It is well known that the image of a man's face lies on an intrinsically three dimensional manifold, which can be described by two pose direction variables and an azimuthal lighting angle. We take the same 698 images each of them being 64 pixel by 64 pixel, that are used in Isomap. There are three alternative kernel functions: a Gaussian kernel $K(x,y)=\exp(-\|x-y\|^2/256)$, a polynomial kernel $K(x,y)=(\langle x,y\rangle+16)^6$ and a sigmoid kernel $K(x,y)=S(\langle x,y\rangle+0.5)$ with $\varepsilon=0.1$ in the first step of the kernel Isomap algorithm.

3 Conclusions

An initial way on how to select a kernel function is proposed. In [8] a method on how to adjust a kernel to improve its performance is brought forward. However, it deals with only one class of kernels. Thus the problem is discussed from a global view to consider all the types of kernels.

References:

[1] SHAWE-TAYLOR J, CRISTIANINI N. Kernel methods for pattern analysis[M]. New York: Cambridge University Press, 2004.

[2] SEBASTIAN H S, DANIEL D. The manifold ways of perception[J]. Science, 2000, 290: 2268-2269.

[3] TENENBAUM J, De SILVA V, LANGFORD J. A global geometric framework for nonlinear dimensionality reduction [J]. Science, 2000, 290: 2319-2323.

[4] ROWEIS S, SAUL L. Nonlinear dimensionality reduction by locally linear embedding[J]. Science, 2000, 290: 2323-2326.

[5] BELKIN M,NIYOGI P. Laplacian eigenmaps for dimensionality reduction and data representation[J]. Neural Computation,2003,15:1373-1396.

[6] ALFRED V A,JOHN E,JEFFERY D U. The design and analysis of computer algorithms[M]. New York:Addison-Wesley,1974.

[7] MARDIA K,KENT J,BIBBY J. Multivariate Analysis[M]. London:Academic Press,1979.

[8] AMARI S,WU S. Improving support vector machine classifiers by modifying kernel functions[J]. Neural Networks,1999,12:783-789.

一个关于如何选择核函数的框架

万海平,何华灿

(北京邮电大学 信息工程学院,北京 100876)

摘　要:如何为特定的机器学习任务选择合适的核函数,是统计学习和核方法理论中的一个具有挑战性的问题。在此从保形映射和流形学习的角度,提出了一种探索性解决方法,并以实验检验这种构想,做出了初步结论。

关键词:核方法;流形学习;保形映射

中图分类号:TP18　　**文献标识码**:A　　**文章编号**:1001-6600(2006)04-0050-04

(责任编辑　马殷华)