International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: publications@iiasa.ac.at
Web: www.iiasa.ac.at

Interim Report IR-08-025

Kernel-based estimation method

Dongling Zhang (zhangdl365@sina.com)

Approved by

Marek Makowski (marek@iiasa.ac.at)
Leader, Integrated Modeling Enviroment Project

October, 2008

Interim Reports on work of the International Institute for Applied Systems Analysis receive only limited review. Views or opinions expressed herein do not necessarily represent those of the Institute, its National Member Organizations, or other organizations supporting the work.

International Institute for Applied Systems Analysis
Registration number: ZVR 524808900

Foreword

This report describes the research of the author advanced during her participation in the 2008 Young Scientists Summer Program (YSSP) with the Integrated Modeling Environment Project. The research deals with data analysis (often also called data mining). Although many data analysis methods exist and are widely used, there are also many data sets for which application of these methods do not provide satisfactory results. Therefore, there is a need for developing new methods that can support knowledge creation from data analysis.

The author has developed a new kernel-based estimation method aimed at predicting relationship between a-set of given variables and one dependent (often called decision or outcome) variable. In many practical applications a functional form of such relationship is unknown, therefore classical regression methods are not applicable. In such situations non-parametric estimation is one of very few approaches that can effectively help in predicting outcomes based on analysis of available data.

The author has proposed a new method and tested it on a set problems, not only those commonly used for testing data mining methods but also on problems researched by fellow YSSPers. The tests have shown better performance of the proposed method than several standard methods of data analysis on several problems. Considering the fact that no data-mining method is good for diversified classes of problems, this is a very good result.

The summer-period of the YSSP is only three months short, and this type of research requires substantially more time for completing. In particular, more work is needed for exploring the potential of data clustering discussed in Section 4. The results presented in this report show that the author not only has already achieved interesting results, but she is also finding a good way to extend them to a form applicable to problems from various application fields that – from analytical point of view – belong to a certain class of data analysis problems.

Abstract

Regression is a basic statistical tool for estimation task of data mining, which is to predict the relationship between a dependent variable and one or more independent variables. Parametric and nonparametric regressions are two kinds of regression approach used for various problems. This work proposes a kernel-based nonparametric regression method, which can solve nonlinear regression problem properly by mapping the data to a higher-dimensional space by kernel function. With this method, we conduct a series of experiment on nonlinear function and real world regression problems, and the results reveal the effectiveness of the model. The results reveal that the model is efficient on some data sets with similar or even higher precision than the prevalently used support vector regression and neural network regression method. Nevertheless, there are still other data sets which kernel-based method cannot works well, such as water flow and forest fire data set.

Keywords: kernel, nonparametric, nonlinear regression, estimation, data mining

Acknowledgments

This report and all the research described here were done during my participation in the Young Scientists Summer Program (YSSP) 2008 at the International Institute of Applied Systems Analysis in Laxenburg, Austria.

I would like to thank Dr. Marek Makowski for his supervision and lots of good advice on my research he has given to me during the program. I also give many thanks to Adam Kiczko, Sayaka Kanata and Shuo Liu for their advices and cooperation.

Finally, I would like to thank Prof. Shi and Prof. Tian from Research Center of Factitious Economy and Data Science of Chinese Academy of Sciences for scientific help and National Natural Science Foundation of China for the financial support which made this participation in the YSSP possible.

About the Author

Dongling Zhang is now a PhD student at the Graduate School of Chinese Academy of Sciences. Her research interests include data mining and knowledge discovery techniques in databases, especially the issues of mathematical programming and optimization methods used for data mining. Dongling's recent research objective is multiple criteria linear programming method used for data classification and regression analysis. At IIASA she was working on a nonlinear regression model based on kernel theory and original multiple criteria linear programming regression method.

Contents

Kernel-based estimation method

Dongling Zhang (zhangdl365@sina.com)*

1 Introduction

1.1 Problem description

With the rapid development of information technology, more and more data are generated by retail companies, governments and all kinds of organizations, even individuals. Today, a lot of companies have realized the importance of data analysis to detect patterns and allow predictions to help making decision. Data mining technologies are developed under these circumstances.

Data mining refers to the analysis of the large quantities of data that are stored in computers to get useful knowledge. Usually, statistical and artificial intelligence analysis tools are used to extract or mine knowledge from large-scale data sets [1, 2]. The extracted knowledge must be new and useful things that have not been known or used before. There are two primary goals of data mining studies, prediction and description. Prediction refers to using some variables in the data set to predict unknown or future values of other variables of interest. Description involves finding some patterns to describe the relationships between data or actions and outcomes.

In this work, I will focus on the prediction task in data mining. Data mining algorithms use the existing data to learn knowledge that can predict the class label of other unclassified data, or predict a continuous value of some attributes. Specifically, real value prediction problem will be the main concerning of my research.

Now, I will use the following example to explain the prediction problem. This prediction task aims at predicting the burned area of forest fires in some region (see Table 1). In this data set, there are 4 input attributes as tempera[*]ture, humidity, wind and rain. And the area attribute is the output or the dependent variable, which has high correlation with the 4 input variables. We notice that in this data set some of the values in the last column are unknown (question mark), which are future values that we want to predict, as fire area detection is the key point for improving firefighting resource management. In order to predict the unknown value, we must extract valuable knowledge from the known data, such as the relationship between input variables and output variable. Usually, the knowledge need to be expressed into the forms of

[*] School of Management, University of Chinese Academy of Sciences, Beijing, China.

decision rules, mathematical formulas, etc. With these rules or formulas, the unknown value can be predicted. For the forest fire burned area prediction, regression method is often applied to solve this kind of problem, which is to predict some real value. Although this forest fire example is much simpler than the real cases, the basic theory of this kind of problem is all the same. From this example, we will be able to know that prediction is such issue that we should use some known data to predict unknown or future values which can help decision making.

Table 1. A simple example for estimation

temperature	**humidity**	**wind**	**rain**	**area**
16.1	44	4	0	49.37
20.1	34	4.5	0	58.3
28.3	26	3.1	0	64.1
8.3	97	4	0.2	0
26.8	38	6.3	0	0.76
19.3	39	3.6	0	1.56
15.4	66	4	0	10.13
21.9	73	7.6	1	0
22.4	54	7.6	0	?
16.1	44	4	0	?
28.3	26	3.1	0	?

To descript the prediction problem more clearly, and to make it more general and consistent with the later part of the report, we illustrate the problem into the formulation with some denotations.

We use data set $T=\{(x_1,y_1),...,(x_l,y_l)\}$ $\in (X \times Y)^l$ to denote all the known data, where $x_i \in X = R^n$ is the input or the independent variable, it has n attributes. $y_i \in Y = R$ is the output or the dependent variable. Different from the classification problem, y_i is not limited to a categorical value, but can be any real number. Data set $T1=\{(v_1,?),...,(v_k,?)\}$, $v_i \in R^n$ is used to denote all the data with unknown value of y, where v_i has the same attributes as x_i. T and $T1$ are homogenous.

The purpose of this kind of problem is to find the relationship between the dependent variable y and independent variable x on the given data set T, so that when given a new input x, we can infer the corresponding y, where y is also a real number. This new input data belongs to data set $T1$ mentioned above.

Usually, regression is the main solution for the estimation problem. Also, many different regression methods have been developed up to now. Each method has its merits and drawbacks, and can be used for particular application.

1.2 Estimation methods in data mining

Regression is the principal and most widely used method for prediction. The key idea of regression is to discover the relationship between the dependent variable and one or more independent variables. So far, many regression approaches have been presented.

Common approach in regression is to fit the data to a global parametric function. In the parametric regression method, the formulation of the global function f, with a series of parameters to be estimated, must be specified in advance. Usually, the regression model fits to the function of $y_i = f(\beta, x_i) + \varepsilon_i,\ (i = 1, \cdots, l)$, where $\beta = (\beta_1, \cdots, \beta_n)'$ is a vector of parameters to be estimated. (x_i, y_i) is the ith pair of training sample of n observations. ε_i is the error which is assumed to be normally and independently distributed with mean 0 and constant variances σ^2.

Least square regression is a most widely used parametric method. It was first developed to fit a straight line by determining the coefficients of each independent variable, which can minimize the sum of squared error over all observations. In multiple regression and some non-linear regression problems, least square can also work well. As a parametric method, for least square regression, we must specify the exact function of the model with some parameters to be estimated. If the relationship between the variables can be modeled by linear function, least square linear regression will be the best choice. But when the linear model is not fit for the data, or when the function is difficult to define in advance, nonparametric regression method is an alternative and efficient way to get a more accurate model.

Nonparametric method is often used for nonlinear regression problem. It estimates the regression function directly rather than to estimate the parameters in the function. A general nonparametric model is written in a similar manner with parametric model, but the function f is left unspecified: $y_i = f(x_i) + \varepsilon_i,\ (i = 1, \cdots, l)$. That is, the function f is to be estimated from the data and no general formulation is given in advance. As to many kinds of real application, it appears to be difficult to know directly the exact relationship between the variables of data set and the formulation of f cannot be specified beforehand. Nonparametric method is very useful for this kind of problem [3].

Some typical nonparametric regression methods are support vector regression, neural network regression, rule-based regression and so on [3].

In this work, I propose a kernel-based nonparametric estimation method, which can be used to solve nonlinear regression problem. In this method, there are two steps, first is to map the data to a higher-dimensional space by kernel function, and then apply linear regression method to fit the new data. Thus provide the solution to nonlinear regression problem.

The outline of the paper is as follows. We start from giving a brief review of kernel method in Section 2. Then Section 3 introduces the kernel-based nonparametric regression method. To demonstrate the effectiveness of the method, the experimental results are provided in Section 4. Finally, the conclusion is given according to the experiment.

2 Kernel theory

Kernel method offers a powerful solution to solve nonlinear separable problem. By projecting the data into a high dimensional feature space, the original nonlinear problem seems to be a linear one. The most well-known application of kernel function is support vector machine (SVM) classifier. To show how the kernel method works, we use the below problem as an example [4].

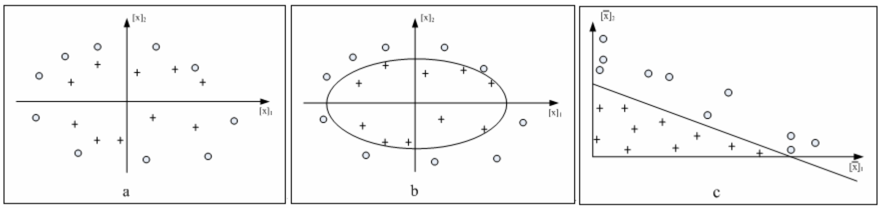


Figure 1. A simple nonlinear classification example

Suppose we are solving the classification problem in Figure 1(a). The classification data set is $C = \{(x_i, y_i),\ i = 1,...,20\}$, where x_i is taken from the space $([x]_1,\ [x]_2)$, and we have $x_i = ([x_i]_1,\ [x_i]_2)^T$, $y_i \in \{-1,1\}$. It is obvious that the best classification line is an ellipse in the space $([x]_1,\ [x]_2)$ shown in figure 1 (b):

$$[w]_1[x]_1^2 + [w]_2[x]_2^2 + b = 0$$

Where $[w]_1$ and $[w]_2$ are coefficients in terms of the input data.

Now, the problem is how to get the two coefficients given the specific data set. On this matter, we notice the fact that if we replace the variable $x_i = ([x_i]_1,\ [x_i]_2)^T$ with $\overline{x_i} = ([\overline{x_i}]_1,\ [\overline{x_i}]_2)^T$ from a new feature space, where $[\overline{x}]_1 = [x]_1{}^2$, $[\overline{x}]_2 = [x]_2{}^2$ then the separation line will be:

$$[w]_1[\overline{x}]_1 + [w]_2[\overline{x}]_2 + b = 0$$

With some linear classification methods, we can get $[w]_1$ and $[w]_2$.

The projection in this problem is:

$$\phi : \begin{array}{l} [\overline{x}]_1 = [x]_1^2 \\ [\overline{x}]_2 = [x]_2^2 \end{array}$$

Where $\Phi : [x] \rightarrow \overline{[x]}$ is a nonlinear map from the input space to some feature space, usually with high dimension. $[x] \in R^2$ is the independent variable with 2 attributes. $\overline{[x]} \in R^2$ is the new independent variable in the high-dimensional space with 2 attributes.

Similar to the above example, the basic way to build a nonlinear classification machine includes two steps: first a projection function transforms the data into a feature space F, and then a linear classification method to classify them in the feature space. On this basis, by projecting the low dimensional data into a space of higher dimension, the data which is nonlinearly separable in the original space, may become linearly separable.

But as is commonly known, the mapping function is always implicit. Thus, if the input data have many attributes, it is hard to perform such mapping operations.

Kernel function offers an alternative way to such projection. We will introduce the kernel methods though SVM classifier.

Suppose a nonlinear separable data set is mapped into the new space by the projection $g: x \to z$, we get the transformed data set $T' = \{(z_1, y_1),\ ...,\ (z_l, y_l)\}$, where $z_i \in R^d$, $y_i \in \{-1,1\}$. In SVM classification, the classifier for data set T' in new space can be formulated by solving the below linear programming problem.

$$\begin{aligned}
&\underset{w,b,\xi}{Minimize} && \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i, \\
&s.t. && y_i (w^T z_i + b) \geq 1 - \xi_i, \\
& && \xi_i \geq 0, \quad i = 1, ..., l.
\end{aligned}$$

Where $z_i = g(x_i) \in R^d, \quad i = 1, ..., l, \quad \xi = (\xi_1, ..., \xi_l)^T$.

To solve this problem, we get its dual model as follows:

$$\begin{aligned}
&\underset{\alpha \in R^l}{Minimize} && \frac{1}{2} \sum_{i=1}^{l} \sum_{i=1}^{l} y_i y_j \alpha_i \alpha_j (z_i \cdot z_j) - \sum_{j=1}^{l} \alpha_j, \\
&s.t. && \sum_{i=1}^{l} y_i \alpha_i = 0, \\
& && 0 < \alpha_i < C, \quad i = 1, ..., l
\end{aligned}$$

Where $z_i = g(x_i) \in R^d$.

We replace $(z_i \cdot z_j)$ with $g(x_i) \cdot g(x_j)$, the above formulation will be:

$$\begin{aligned}
&\underset{\alpha \in R^l}{Minimize} && \frac{1}{2} \sum_{i=1}^{l} \sum_{i=1}^{l} y_i y_j \alpha_i \alpha_j (g(x_i) \cdot g(x_j)) - \sum_{j=1}^{l} \alpha_j, \\
&s.t. && \sum_{i=1}^{l} y_i \alpha_i = 0, \\
& && 0 < \alpha_i < C, \quad i = 1, ..., l
\end{aligned}$$

If we can compute the inner product $(g(x_i) \cdot g(x_j))$ in feature space directly, there is no need to get the mapping function $g: x \to z$. And by this means, we can also combine the two steps of generating a nonlinear classification machine to one process. To this end, the kernel function is defined to be such a direct computation method.

Definition of kernel function:

Function K is a kernel if for all $x_i, x_j \in X$, $K(x_i, x_j) = g(x_i) \cdot g(x_j)$, here g is a mapping from X to an inner product feature space. Although g is always implicit, kernel function K can be explicit. Some commonly used kernel functions are polynomial kernels, Gaussian RBF kernels and Mercer kernels. The function of the RBF kernels is $K(x_i, x_j) = \exp(-\| x_i - x_j \|)^2 / \sigma^2)$.

In the formulation of some methods, such as SVM, the input data appear to be included in the inner product formula, so it is convenient to use kernel function K rather than mapping function g. After putting kernel function K into the SVM formulation, we can get the optimization result which consists of a series of value of α. With these values, we can obtain the classifier for the nonlinear separable data set.

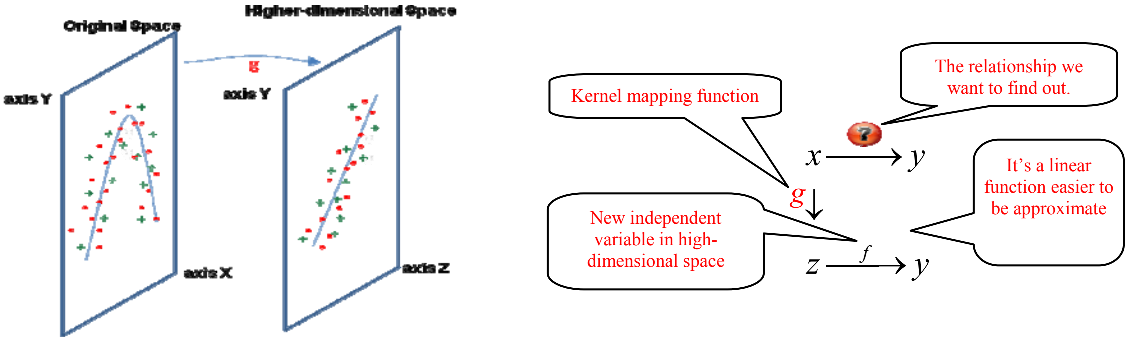
To sum up, kernel method is a powerful way to solve nonlinear separation problem. With kernel function, we don't need to know the exact form of the mapping function and can easily implement the projection.

3 The proposed kernel-based estimation method

Let's consider the regression problem firstly. The training set of regression problem is denoted by $T = \{(x_1, y_1), ..., (x_l, y_l)\} \in (X \times Y)^l$, where $x_i \in X = R^n$ is the input or the independent variable, it has n attributes. $y_i \in Y = R$ is the output or the dependent variable. Different from the classification problem, y_i is not limited to a categorical value, but can be any real number.

From Section 2, we know that for nonlinear separable data, when mapped to a higher dimensional space by kernel method, the data seem to have more chance to be linear separable. After that, we can apply any linear classification method to it, which will be much easier than solve the nonlinear separable problem directly.

Similarly, for a regression problem, if we can map the data to a higher dimensional space by kernel method, the nonlinear regression problem might be solved by linear regression method properly. The principle of this idea can be shown in the Figure 2.


Original Space
Higher-dimensional Space
g
axis Y
axis Y
axis X
axis Z
Kernel mapping function
The relationship we want to find out.
?
x
y
g
New independent variable in high-dimensional space
z
f
y
It's a linear function easier to be approximate

(a)

(b)

Figure 2. The principle of Kernel-based estimation method

In nonlinear classification problem, kernel function, usually taking the form of inner product of two samples, is an alternative way to perform the projection instead of mapping a single data with an explicit mapping function. In SVM, the data appear to be included in the inner product formulation, so it is easy to be mapped with kernel function. But for regression problem, there is no such kind of formula as SVM, the main problem will be how to map the data if there is only kernel function instead of explicit mapping function. The solution of this problem is similar with random projection method [5, 6]. But we put all the training data to the projection function instead of random sample.

Kernel mapping function is the main part of this kernel-based nonparametric regression method. Suppose x is the data to be mapped, and the mapping function g is as follows.

$g(x) = (K(x, x_1), ..., K(x, x_l))$,

Where $K(x, x_i)$ is an arbitrary kernel function. Here we use Gaussian RBF kernel.

By this process, each data can be mapped to a *l*-dimensional space, and the transformed data set in this new space is $T' = \{(z_1, y_1), ..., (z_l, y_l)\}$, where $z_i = g(x_i) \in R^l$. For this new data set, linear regression method can be used, and then get the relationship between input x and output y. In the linear regression part of this method, there are many choices, such as least square regression, MCLP for regression [7] and so on. In this paper, we choose the least square as the linear regression criteria to perform the experiment.

In the linear regression part, we need to fit the model $y_i = f(\beta, z_i) + \varepsilon_i,\ (i = 1, \cdots, l)$, where $\beta = (\beta_1, \cdots, \beta_l)'$ is a vector of parameters to be estimated. (z_i, y_i) is the i th pare of training sample of n observations. ε_i is the error which is assumed to be normally and independently distributed with mean 0 and constant variances σ^2. After β is estimated, the mapping from *z* to y $f(z): z \to y$ is figured out.

The procedure of kernel-based estimation method is shown in Figure 3.

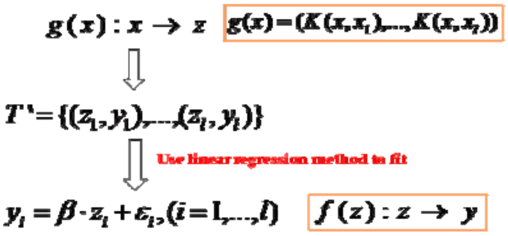

$g(x): x \to z$
$g(x)=(K(x,x_1),...,K(x,x_l))$
$T'=\{(z_1,y_1),...,(z_l,y_l)\}$
Use linear regression method to fit
$y_i = \beta \cdot z_i + \varepsilon_i, (i=1,...,l)$
$f(z): z \to y$

Figure 3. The procedure of Kernel-based estimation method

For a new input variable, in order to predict its corresponding *y*, such a process is also needed to be performed. At first, use mapping function g(x) to project the input data into z, which is in higher-dimensional space. Secondly, apply estimation function f(z) to predicting the value of y, where $y = f(z) = \beta \cdot z$, β and *z* are all *l*-dimensional vectors.

Actually, the two steps can be combined into one after β is estimated. For a new input variable x, in terms of the above analysis, in the first step we have $z = g(x) = (k(x, x_1), ..., k(x, x_l))$. For the second step, if we expand the $\beta \cdot z$ with the components of vector β and *z*, it will be $\beta_1 * K(x, x_1) + ... + \beta_l * K(x, x_l)$. That is, we can use the function $y = \beta_1 * K(x, x_1) + ... + \beta_l * K(x, x_l)$ to predict the corresponding y of each new input x.

4 Results of experiment

To prove the effectiveness of this method, we now apply it to some data sets, including one nonlinear function data set, three UCI regression data sets and some real application data sets. And, in order to evaluate our method, experiments are also conducted with other regression methods like multiple regression, support vector regression and neural network regression. In the experiments, we use 10-folds cross-validation method to get the precision of each method. Mean standard error (MSE) and mean absolute deviation (MAD) are the major criteria to evaluate the results. For each method, grid search approach is used to choose the best parameters with highest precision for one certain data set. Here we list some of the results.

Some evaluation criteria listed in the form are calculated by the following formulation.

Mean Squared Error: $MSE = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

Mean Absolute Deviation: $MAD = \frac{\sum_{i=1}^{n}|y_i - \hat{y}_i|}{n}$

Max Deviation: MaxD = max($|y_i - \hat{y}_i|$), for i=1,…,l

Min Deviation: MinD = min($|y_i - \hat{y}_i|$), for i=1,…,l

Mean Predictive Error: $MPE = \frac{\sum_{i=1}^{n}|\frac{y_i - \hat{y}_i}{y_i}|}{n} * 100\%$

Max Predictive Error: MaxPE = max($|\frac{y_i - \hat{y}_i}{y_i}| * 100\%$), for i=1,…,l

Min Predictive Error: MinPE= min($|\frac{y_i - \hat{y}_i}{y_i}| * 100\%$), for i=1,…,l

Predictive error: $|\frac{y_i - \hat{y}_i}{y_i}| * 100\%$

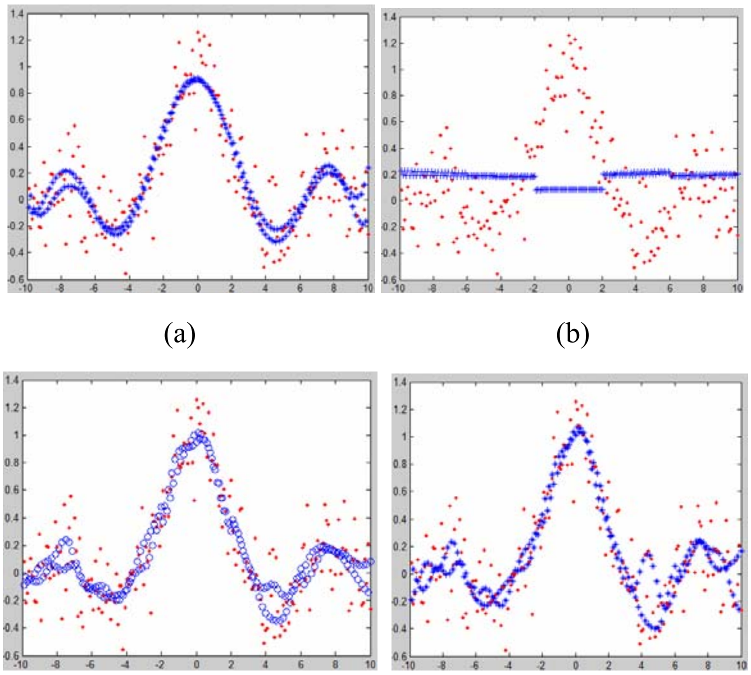
PE10: Predictive error within 10%

PE20: Predictive error within 20%

4.1 Data from nonlinear function sinc

The training samples are based on 200 points drawn from the function sinc(x)=sin(x)/x, where a random normal noise is added to the value of y.

The following Figure shows the results of fitting the data with four methods: (a) kernel-based nonparametric regression (Kernel), (b) multiple linear regression (MR), (c) neural network regression (NN), (d) support vector regression (SVR).



(a) (b)

(c) (d)

Figure 3. Nonlinear function regression results

Table 2. The result of function approximation

	Kernel	MR	NN	SVR
parameter	$\sigma^2 = 1000$		Spread=0.5	g =1 ε =0.1
MSE	**0.0437**	0.191	0.0446	0.0501
MAD	**0.1643**	0.3472	0.1659	0.177

Figure 3 shows that the kernel-based method seems to be more smoother than the other methods and in this data set it has the best fit result. Also, we will be able to know that there must be rather bad result if directly use multiple linear regression method to fit nonlinear function. Meanwhile, in Table 2 we know kernel method has the best predictive precision with the lowest MSE and MAD value.

4.2 Real cases

In dealing with real word cases, we employ the proposed approach to several UCI and StatLib data sets [8, 9], including forest fires, housing and space_ga data sets. We also apply the method to Poland Narew river water flow data set and crop output data set of China. Some of the results are listed here.

4.2.1 UCI Housing data set

This data set is taken from UCI repository, which concerns housing values in suburbs of Boston. It contains input attributes relevant with crime rate, proportion of residential land, distance to employment centers, and accessibility to radial highways, etc. The output variable is median value of owner-occupied homes in $1000's. Some quantitative characteristic of data are listed in Table 3 Before analysis, the input variables are scaled to the range of [-1,1].

Table 3. Quantitative characteristic of data

Sample Size	506
The number of input attributes	13
Range of output value	[5,50]
Scaling range of input X	[-1,1]

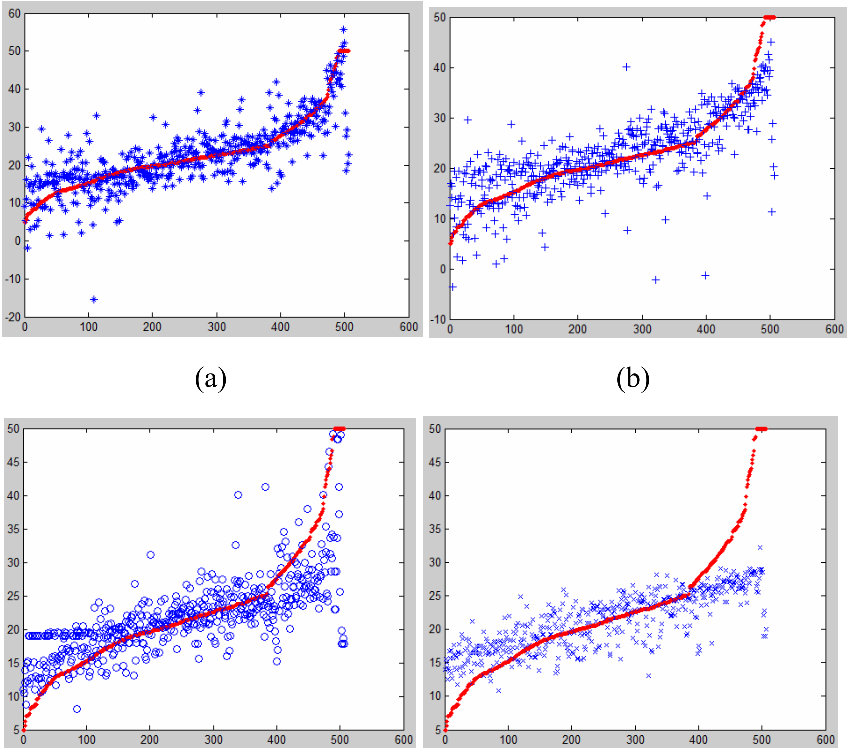
Table 4 shows the results of estimation with 4 different methods. From the two rows of predictive error within 10%(PE10) and 20%(PE20), we can see each method can achieve satisfiable precision on most of the samples. This can also be verified by Figure 4, which shows both the observation value (red curve) and predictive value (blue point) of the output variable of each sample. The horizontal axis is the number of samples, and the vertical axis is the value of output. To make the figure clear, we sort the samples in the figure by ascending their observation value of output.

Compared with other methods, kernel method has the lowest predictive error which can be shown by the criteria MSE and MAD.

Table 4. Evaluation criteria and parameters of methods

	Kernel	MR	NN	SVR
parameter	$\sigma^2 = 10^6$		Spread=0.25	g =0.1 ε =0.5
MSE	32.913	36.264	40.619	44.912

MAD	3.8891	4.1036	4.1607	4.4231
MaxD	31.5190	38.7437	32.2	31.0546
MinD	0.0057	0.0093	0.023	0.0029
MPE	21.1%	21.86%	20.3%	21.02%
MaxPE	215.6%	237%	163.9%	211.4%
MinPE	0.03%	0.027%	0.1%	0.014%
PE10	211(41.7%)	199(39.3%)	208(41.1%)	188(37.2%)
PE20	339(67%)	332(65.6%)	331(65.4%)	323(63.8%)



(a) (b)

(c) (d)

Figure 4. UCI housing data estimation results

4.2.2 StatLib Space_ga data set

This data set is taken from StatLib, which concerns the proportion of votes cast for both candidates in the 1980 presidential election. It contains 3,107 observations on U.S. county votes cast in the 1980 presidential election. The input attributes include the population, education, housing, income and spatial information of each county. The output variable is the log of the proportion of votes cast for both candidates. Some additional information of data set are listed in Table 5. Before analysis, the input variables are scaled to the range of [-1,1].

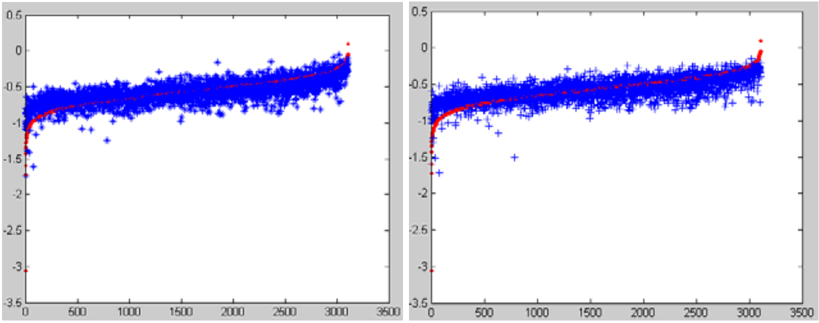
Table 5. Quantitative characteristic of data

Sample Size	3107
The number of input attributes	6
Range of output value	[-3.0570 , 0.1001]
Scaling range of input X	[-1,1]

Table 6 and Figure 5 show the results of estimation with 4 different methods. Similar with Figure 4, the samples in Figure 5 are also arranged by ascending the observation value of output. Both characters in Table 6 and Figure 5 can show the considerable precision of each method. And obviously, kernel method has the lowest predictive error.

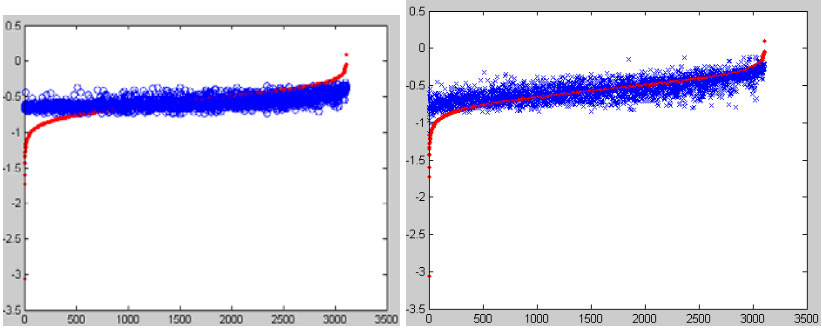
Table 6. Evaluation criterion and parameters of methods

	Kernel	MR	NN	SVR
parameter	$\sigma^2 = 10^6$		Spread=0.5	g =0.1 ε =0.05
MSE	0.015	0.0178	0.0257	0.0176
MAD	0.093	0.1001	0.1218	0.0974
MaxD	1.2438	1.4578	2.4349	1.9146
MinD	2.9806e-006	1.081e-004	5.2730e-005	1.7506e-006
MPE	18.65%	20.53%	25.87%	19.03%
MaxPE	411.1%	794.9%	808.67%	513.4%
MinPE	0.0005%	0.018%	0.009%	0.0004%
PE10	1185 (38.14%)	1128(36.3%)	887 (28.5%)	1165(37.5%)
PE20	2089 (67.24%)	2050 (66%)	1696 (54.6%)	2055 (66.1%)



(a)

(b)



(c) (d)

Figure 5. StatLib Space_ga data estimation results

4.2.3 UCI Forest fires data set

This data set is taken from UCI repository, where the aim is to predict the burned area of forest fires. The input attributes include some weather conditions, a few fire indexes, etc. The output variable is the burned area of the forest. Some quantitative characteristic of data are listed in Table 7. Before analysis, the input variables are scaled to the range of [-1,1].

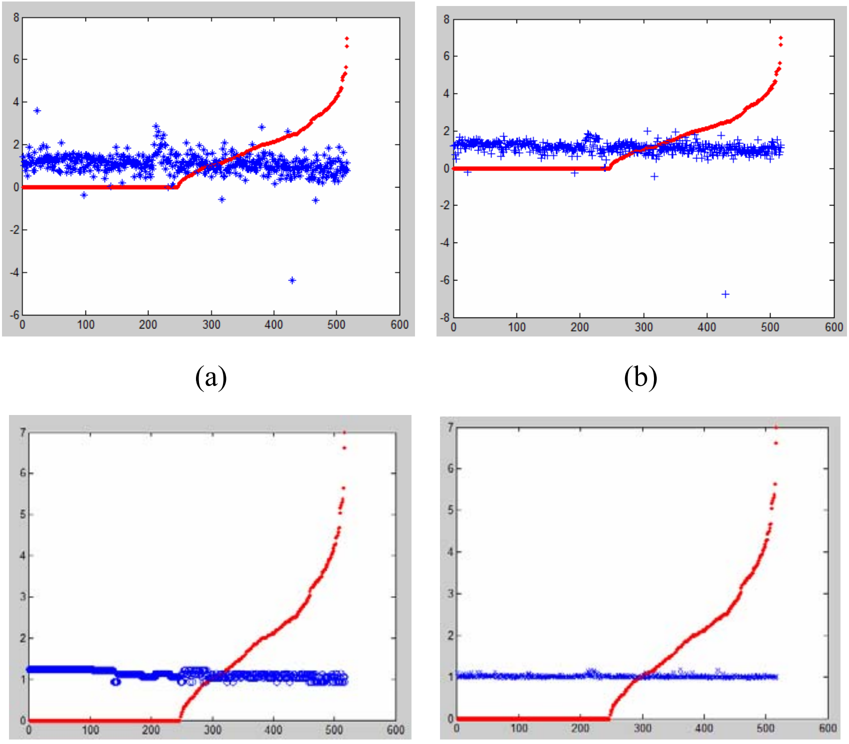
Table 7. Quantitative characteristic of data

Sample Size	517
The number of input attributes	8
Range of output value	[0,6.9956]
Scaling range of input X	[-1,1]

Table 8 and Figure 6 show the results of estimation with 4 different methods. Similar with Figure 4, the samples in Figure 6 are also arranged by ascending the observation value of output. From Figure 6, it seems that no method can work well on this data set. On the highest part of the observation value, the predictive precision is much lower than the middle part.

Table 8. Evaluation criterion and parameters of methods

	Kernel	MR	NN	SVR
parameter	$\sigma^2 = 10^{13}$		Spread=0.7	g =0.1 ε =1
MSE	2.1662	2.3784	2.1101	1.9776
MAD	1.2225	1.2489	1.2065	1.1435
MaxD	6.8662	9.2211	6.0590	6.0012
MinD	5.8412e-004	0.0038	8.0422e-007	0.0046



(a) (b)

(c) (d)

Figure 6. UCI Forest fires data estimation results

In addition to the above data sets, experiments on other data sets have also been conducted, such as Poland Narew river water flow data set, crop output data set in one region of China, etc. The results on different data sets demonstrate that the performance of kernel method varies with data set.

4.3 Analysis of results

The above experiments on some data sets indicate that the kernel-based estimation method is efficient on some data sets with similar or even higher precision than the prevalently used support vector regression and neural network regression methods. Nevertheless, there are still other data sets which kernel-based method does not work, such as water flow and forest fire data set. For this kind of data sets, we are trying to find out the reasons of unsatisfactory performance in order to improve the performance of our method. We sum up some here the possible reasons, which might influence the result of estimation.

Of the possible reasons, the quality of data and the characteristic of data are important factors to the final data analysis result. For example, if the variables are correlated, regression-based techniques are unreliable and this will give misleading output. Therefore, such data preprocessing techniques as cluster, principal component analysis and attribute selection may be necessary for a certain kind of data set.

4.3.1 Cluster of data

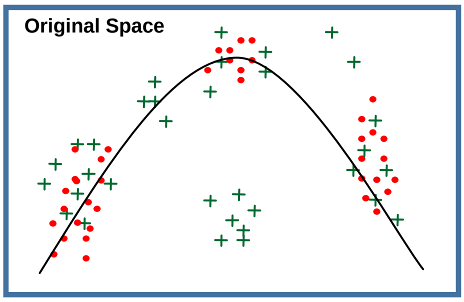


Figure 7. Clusters of data

Figure 7 shows the clusters in one data set. The red points are training samples, and green cross points are test samples.

As is shown in Figure 7, the samples may locate in different subsets. If we try to find a universal function for the whole data set, it may not work well. In this case, the best way is to detect the different subsets (clusters) in the whole data set, then construct a different model for each cluster.

The second reason to cluster the data is the training sample and test sample may be far away from each other, or they may be in different clusters. So using a model constructed in the training sample to predict the test sample may not provide good predictions. This conclusion has been validated by the experiment on 2-dimensional data set.

So if we can construct an individual model for each subset separately, then the result might be improved. But how to detect the clusters best suitable for the following estimation process in the data set is the main problem.

After the preliminary cluster experiment with k-means method on the forest fires data set (over 3 dimensions), we did not find any good measure to detect clusters that can improve the predictive result.

Definitely, instead of Euclidean distance as the similarity measure of data in k-means method, there are other methods and similarity measure can be used to detect clusters. Examination of various methods for finding characteristics of data and suitable similarity measures will be the future work of us.

4.3.2 Principal component analysis

Since real data set may include a lot of variables which might be correlated. And these patterns in data are hard to be identified in high-dimensional data set, where graphical representation is not available. If we use such correlated variables to build predictive models, the models might be misleading. Therefore, finding patterns in the data and revealing the internal structure of the data in a way which best explains the variance in the data will be the main task in data preprocessing of some applications.

Principal component analysis (PCA) is a powerful tool for such kind of data analysis. PCA can highlight the similarities and differences of variables. It can compress the data, by reducing the number of dimensions, without much loss of information. By this way, the data set can retain those characteristics which contribute most to its variance. The retained characteristics are so-called principal components. Using them in building predictive models may improve the reliability and the predictive precision of models.

However, such analysis is usually application dependent, and may not always provide good results.

4.3.3 Attribute selection

In real cases, there always are many attributes in the data set. For example, in the forest fires data set, four kinds of attributes are included in the input variables, spatial, temporal, weather and fire weather index variables. But only some of them are relevant to the forest fires area, and it is better to use these variables instead of all to build predictive model. Because additional attributes can interfere with other more useful attributes, prediction with additional attributes can affect the result. Put many noisy attributes into an algorithm, it will be difficult for the algorithm to separate signal from noise. The more noisy attributes available, the less likely the algorithm will be able to properly identify those attributes that lead to good generalization given a fixed training sample [10]. Therefore, instead of building the predictive model with all the attributes, attribute selection need to be done for some data sets.

But, how to select the best attributes to use for predictive task is also a main problem here. Simply eliminating irrelevant attributes will not help. For this purpose, several attribute selection algorithms are available to fulfill this process, such as randomized algorithm and sequential search algorithm. For some data sets with low predictive precision, probably attribute selection before building predictive model might be the remedy.

5 Summary

This work proposes a kernel-based nonparametric regression method, which can solve nonlinear regression problem properly by mapping the data into a higher-dimensional space by kernel functions. With this method, we conduct a series of experiment on nonlinear function and real world regression problems, and the results reveal the effectiveness of the model. The results demonstrate that the approach is efficient on some data sets with similar or even higher precision than the prevalently used support vector regression and neural network regression methods. Nevertheless, there are still other data sets for which kernel-based method cannot works well, such as water flow and forest fire data set.

Actually, no method can perform well on all sorts of problems; therefore there will be many problems that require further research. Some of them, for instance, are: (i) using clustering techniques and choosing appropriate similarity measures to detect the subset of data; (ii) finding out the characteristic of data on which kernel method can predict well or not, etc.

References

[1] Y. Shi. Data mining. In: M. Zeleny (Ed.), IEBM Handbook of Information Technology in Business,pp. 490–495. International Thomson Publishing, England. (2002).

[2] J. Han and M. Kamber. Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, California. (2001).

[3] J. Fox, “Multiple and Generalized Nonparametric Regression”, Sage university papers series on Quantitative Applications in the Social Sciences, CA: Sage. Thousand Oaks, 2000, pp.07-131

[4] N.Y. Deng and Y.J. Tian, New Approach in Data Mining – Support Vector Machine, Science Press, Beijing, 2004

[5] D. Fradkin and D. Madigan, “Experiments with random projections for machine learning”, in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2003, pp. 571–522

[6] M.F. Balcan, A. Blum and S. Vempala, “Kernels as features: On kernels, margins, and low-dimensional mappings”, Machine Learning 65(1), 2006, pp.79-94

[7] D. Zhang, Y.J. Tian, and Y. Shi, “A Regression Method by MCLP”, Proceeding of Conference on Multi-criteria Decision Making , 2008. (Working Paper).

[8] Asuncion, A. and Newman, D.J. UCI Machine Learning Repository [http://www.ics.uci.edu/~mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science. 2007

[9] StatLib archive of statistical data sets and software, [http://lib.stat.cmu.edu/modules.php], Pittsburgh, PA: Carnegie Mellon University, the department of statistics

[10] Caruana, R., and Freitag, D., “Greedy attribute selection”, In Proc. ML-94. Morgan Kaufmann (1994)