Incorporating Method of Time Correlation and SVM Based on Time Geodesic Distance

Ping WANG\textsuperscript{a,b}, Guisheng ZHANG\textsuperscript{a} \textsuperscript{*}

\textsuperscript{*}Shanxi University, Taiyuan, 030006, P. R. China
\textsuperscript{b}Shanxi Police Academy, Taiyuan, 030021, P. R. China

Abstract

Support vector machine (SVM) has been used in many fields as a new learning method developed in recent years. When dealing with time series forecasting problem one encounters time correlation prior knowledge of time series data. If prior knowledge at hand can be incorporated into Support Vector learning machines, the generalization performance of SVM may be improved efficiently. In order to incorporate time correlation into SVM, this paper presents time geodesic distance for structural feature of learning data, and this presented new metric can be made use of by classification methods based on distance in training of learning machine. Comparing with the traditional SVM based on air quality database, the presented approach can greatly improve the generalization performance of SVM.

© 2011 Published by Elsevier Ltd. Open access under CC BY-NC-ND license.
Selection and/or peer-review under responsibility of ICAE2011.
Keywords: Support Vector Machine; Time Series Forecasting; Time Geodesic Distance

1. Introduction

SVM introduced by Vapnik [1] is a new promising pattern classification technique, which can obtain good prediction performance and achieve global optimization solution simultaneously. Due to being based on statistical learning theory, SVM can overcome the curse of dimensionality and over-fitting problem which traditional method can’t avoid. According to statistical learning theory, SVM applies the structural risk minimization principle, which seeks to minimize an upper bound of generalization error rather than minimizing training error as the traditional learning machine. At present, SVM has been raised as a power tool for solving numerical pattern recognition [2], face detection [3], text categorization [4] and protein fold recognition [5]. In time series data prior knowledge about time correlation is very useful in time series forecasting problems, however, support vector machine doesn’t take into account this prior knowledge which can improve obviously generalization performance of SVM by using the appropriate

* Corresponding author. Tel.: +0-133-334-37266
E-mail address: wp2004@sxu.edu.cn
method of incorporating time correlation. However, SVMs assume that data is independently and identically distributed (iid), which is not appropriate for tasks such as time series forecasting. In particular, SVMs cannot consider dependencies in the time correlation of adjacent training data. In this paper, we present a novel metric definition based on correlation of time series data, which can efficiently improve the performance of SVM, incorporating time correlation among adjacent data.

2. Time geodesic distance

SVM assume that the individual instances are iid. This is not appropriate if the instances present latent correlation on time feature. That is, it is reasonable a training instance will have a similar label with its neighbors. Although there are some standard machine learners, such as Naive Bayes, logistic regression (LR), and support vector machines (SVMs), produce effective performances in many domains[6], these methods don’t take the knowledge of learning task into account. SVM employs a distance function to determine how close an input vector is to each stored data, but, many learning data lie in dependences on time correlation. Especially for those data points which is time series data, Euclidean distance can not reflect the real distance between two points[7]. In 2000, geodesic distance was proposed by Tenenbaum[8]. The basic idea is that for a neighborhood of points on a manifold the Euclidean distance provide a fair approximation of geodesic distance. For faraway point the geodesic distance is estimated by the length of the shortest path through neighboring points. We present a new metric distance, time geodesic distance, based on time correlation and geodesic distance, and apply it to support vector machine by substituting an estimated time geodesic distance for the conventional Euclidean distance. These techniques are able to represent complex dependencies among data instances on time correlation, giving learning machine higher accuracy on time series forecasting task by incorporating time correlation into algorithm.

Definition 1. Time geodesic distance (TGD): Time geodesic distance between two points \( p \) and \( q \) is the length of the shortest path from \( p \) to \( q \) in chronological order. Suppose \( P = \{p_1, p_2, ..., p_n\} \) is a path between points \( p_1 \) and \( p_n \), where \( p_i \) and \( p_{i+1} \) are connected neighbors for \( i \in \{1,2,...,n-1\} \) and \( p_i \) belong to the time domain for all \( i \). The path length \( d_{TGD}(P) \) is defined as \( d_{TGD}(P) = \sum_{i=1}^{n-1} d_i(p_i, p_{i+1}) \) the sum of the neighbor distances \( d_i(p_i, p_{i+1}) = \|p_i - p_{i+1}\| \) from time domain \( i \) to time domain \( i+1 \).

On the basis of above analysis, the presented approach for training SVM based on time geodesic distance can be concluded as follows:

Step1: Construct neighborhood graph: Assuming nodes \( i \) and \( j \) correspond to \( x_i \) and \( x_j \) on the graph, and they are connected by an edge if they are closer than \( \epsilon \) in time, or if \( x_i \) is one of the k nearest neighbors of \( x_j \) in time. Set edge lengths equal to \( d_{TGD}(i, j) = \begin{cases} d(x_i, x_j) & \text{if } i \in z(x_j) \text{ or } j \in z(x_i) \\ \infty & \text{if } x_i \text{ and } x_j \text{ are not neighbored} \end{cases} \), where \( z(x) = \{x \mid x \in X \text{ is neighbor of } x \text{ in time} \} \).

Step2: Compute shortest paths: For each value of \( k = 1,2,...,n \) replace all entities \( d_{TGD}(i, j) \) by \( \min \{d_{TGD}(i, j), d_{TGD}(i, k) + d_{TGD}(k, j) \} \). The matrix of final values \( D_{TGD} = \{d_{TGD}(i, j)\} \) will contain the shortest path distance between all pairs of points in neighborhood graph.

Step3: Train SVM on time geodesic distance: Let time geodesic distance replace Euclidean distance in support vector machine, then train SVM on training data set.

Here we assume that time series data lie on underlying manifold in time, and points far apart in time can be measured by their shortest path, which can reflect the true time correlation of time series data. Time geodesic distance based on time correlation of time series data can reflect the real distance of two data points in chronological order, and it is superior to the traditional Euclidean distance for structural data set in time. A possible interpretation of Kernel function effects is that they represent dot products in
some feature space, i.e., \( k(x_i, x_j) = \phi(x_i) \times \phi(x_j) \), where \( \phi \) is a map from input space \( X \) into \( F \).

**Definition 2.** Time geodesic distance kernel function: Given a function \( K: X \times X \to \mathbb{R}^N \) and \( x_1, \ldots, x_n \in X \), the kernel matrix or Gram matrix \( \Omega \in \mathbb{R}^{N \times N} \) (where \( \Omega_{ij}=K(x_i,x_j) \)) is positive definite, where \( K(\|x_i - x_j\|) = K(d_{tg}(x_i, x_j)) \) and \( d_{tg}(x_i, x_j) \) is the time geodesic distance of point \( x_i \) and point \( x_j \).

### 3. Results and discussions

Prediction of air quality has been a popular and important focus for atmospheric and environmental research today. This paper presents a new time geodesic distance based on time correlation for structural data and does the simulation experiments on air pollutant database in Hong Kong downtown. Air quality is mostly determined by several pollutants, i.e., carbon monoxide (CO), nitric oxide (NO), nitrogen dioxides (NO2), sulphur dioxide (SO2), nitrogen oxides (NOx), and respirable suspended particulate (RSP). The air quality database used in the simulation experiments was measured at Causeway Bay roadside monitoring station in 2001. All data are provided by the Hong Kong Environmental Protection Department (HKEDP), and the aim of experiments is to compare the generalization performance of presented SVM based on time geodesic distance(TGDSVM) with the corresponding traditional SVM by predicting respirable suspended particulate (RSP) levels in downtown area of Hong Kong during 2001. In simulation, the data of the first ten days in January, 2001 are selected as training sets, which consists of 240 training examples, and the next 24 points serve as test examples corresponding to the hourly measurements on the 11th day of selected month. Time geodesic distance kernel function constructed was base on Gaussian kernel function(RBF), then we can obtain the new time geodesic distance Gaussian kernel function(TGDRBF, \( K_{tg}(x_i, x_j) = \exp(-d_{tg}^2 / \sigma^2) \)). In task, the generalization performance of TGDRBF was evaluated by comparing with the performance of RBF, which was used as a benchmark kernel function. The parameters about program are \( C=\infty \), \( \sigma=80 \), and \( \epsilon=0.001 \), respectively. In both models, the input variables include five pollutants: SO2, NOx, NO, NO2, CO, while outputs are RSP.

#### 3.1. Recovery performances of TGDSVM and SVM

In order to evaluate the recovery performance of TGDSVM, the simulation experiments are carried out. In the experiments, 24 data of January 2001 are selected as learning data set. The experiments compare the recovery performance between the TGDSVM and SVM on training set. Table 1 summarizes the experiment results of two approaches with same parameter settings. Fig.1 illustrates the respective results of original value, SVM and TGDSVM. In general, both models show good recovery performance on the training data. But TGDSVM can deal with the individual data better than SVM.

<table>
<thead>
<tr>
<th></th>
<th>MAE</th>
<th>RMSE</th>
<th>WIA</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SVM</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=180</td>
<td>20.1050</td>
<td>25.5194</td>
<td>0.8587</td>
</tr>
<tr>
<td><strong>TGDSVM</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=180</td>
<td>19.6048</td>
<td>25.3351</td>
<td>0.8639</td>
</tr>
<tr>
<td>k=200</td>
<td>19.8238</td>
<td>25.3264</td>
<td>0.8675</td>
</tr>
</tbody>
</table>
3.2. Predictions of RSP levels

As a real world experiment, we tried to incorporate time correlation for prediction of air quality task. The experiments in this paper compare the RSP levels predicted by TGDSVM and SVM for reviewing the prediction performance of both learning machine models. In order to evaluate the performance of TGDSVM, many groups of experiments are carried out. Figure 2 shows the comparison of TGDSVM and SVM for different values of k. From those figures, it can be seen that the difference between GDSVM and SVM is much larger than in the recovery performance experiment. Obviously, TGDVM has the best prediction performance better than SVM. Table 2 presents prediction errors of two algorithms for testing data in January. It can be seen that the prediction results of TGDSVM are closer to the original data than that of SVM in different parameter k. In other words, the TGDSVM expresses better generalization performance. According to table 2, the best performance of TGDSVM has been obtained for parameter k=190. What is noteworthy in these experiments is that our proposed method is much better than the standard SVM.

Table 2 Prediction Errors of Testing Time Series.

<table>
<thead>
<tr>
<th></th>
<th>MAE</th>
<th>RMSE</th>
<th>WIA</th>
</tr>
</thead>
<tbody>
<tr>
<td>SVM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=170</td>
<td>77.7950</td>
<td>83.4242</td>
<td>0.3222</td>
</tr>
<tr>
<td>k=180</td>
<td>75.7579</td>
<td>85.5418</td>
<td>0.2870</td>
</tr>
<tr>
<td>k=190</td>
<td>68.9000</td>
<td>78.6887</td>
<td>0.3161</td>
</tr>
<tr>
<td>k=200</td>
<td>72.0066</td>
<td>80.1831</td>
<td>0.3162</td>
</tr>
<tr>
<td>TGDSVM</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=170</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=180</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=190</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>k=200</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
4. Conclusions and future work

This paper presents time series kernel function based on time correlation, which can help to improve
the generalization performance of learning machine. We have shown results that are superior to the
standard SVM. The main contribution of TGDSVM is incorporating the time correlation feature of time
series into SVM. Simulating results have proven the effectiveness of the proposed learning machine.
Comparing with traditional SVM, the TGDSVM is much better to predict the time series data, therefore it
appears promising to try the new distance metric on the task. However, it may be worthy to test its value
in more areas.

References

  gaussian kernels to radial basis function classifiers (A.I. Memo No.1599, Massachusetts Institute Of Technology, Artificial
  Intelligence Laboratory, 1996).
[5] C. Ding, I. Dubchak, Multi-class protein fold recognition using support vector machines and neural networks, Bioinformatics
[8] Tenenbaum, J. B., de Silva, V., & Langford, J. C. A global geometric framework for nonlinear dimensionality reduction,