Hindawi Publishing Corporation
The Scientific World Journal
Volume 2014, Article ID 851017, 8 pages
http://dx.doi.org/10.1155/2014/851017

Research Article

Multivariate Time Series Similarity Searching

Jimin Wang, Yuelong Zhu, Shijin Li, Dingsheng Wan, and Pengcheng Zhang

College of Computer & Information, Hohai University, Nanjing 210098, China

Correspondence should be addressed to Jimin Wang; wangjimin@hhu.edu.cn

Received 21 February 2014; Revised 4 April 2014; Accepted 13 April 2014; Published 8 May 2014

Academic Editor: Jaesoo Yoo

Multivariate time series (MTS) datasets are very common in various financial, multimedia, and hydrological fields. In this paper, a dimension-combination method is proposed to search similar sequences for MTS. Firstly, the similarity of single-dimension series is calculated; then the overall similarity of the MTS is obtained by synthesizing each of the single-dimension similarity based on weighted BORDA voting method. The dimension-combination method could use the existing similarity searching method. Several experiments, which used the classification accuracy as a measure, were performed on six datasets from the UCI KDD Archive to validate the method. The results show the advantage of the approach compared to the traditional similarity measures, such as Euclidean distance (ED), cynamic time warping (DTW), point distribution (PD), PCA similarity factor (S_{PCA}), and extended Frobenius norm (Eros), for MTS datasets in some ways. Our experiments also demonstrate that no measure can fit all datasets, and the proposed measure is a choice for similarity searches.

1. Introduction

With the improving requirement of industries for information and the rapid development of the information technology, there are more and more datasets obtained and stored in the form of multidimensional time series, such as hydrology, finance, medicine, and multimedia. In hydrology, water level, flow, evaporation, and precipitation are monitored for hydrological forecasting. In finance, stock price information, which generally includes opening price, average price, trading volume, and closing price, is used to forecast stock market trends. In medicine, electroencephalogram (EEG) from 64 electrodes placed on the scalp is measured to examine the correlation of genetic predisposition to alcoholism [1]. In multimedia, for speech recognition, the Australian sign language (AUSLAN) is gathered from 22 sensors on the hands (gloves) of a native Australian speaker using high-quality position trackers and instrumented gloves [2].

A time series is a series of observations, $x_i(t)$; $[i = 1, \ldots, n; t = 1, \ldots, m]$, made sequentially through time where i indexes the measurements made at each time point t [3]. It is called a univariate time series when n is equal to 1 and a multivariate time series (MTS) when n is equal to or greater than 2.

Univariate time series similarity searches have been broadly explored and the research mainly focuses on representation, indexing, and similarity measure [4]. A univariate time series is often regarded as a point in multidimensional space, so one of the goals of time series representation is to reduce the dimensions (i.e., the number of data points) because of the curse of dimensionality. Many approaches are used to extract the pattern, which contains the main characteristics of original time series, to represent the original time series. Piecewise linear representation (PLA) [5, 6], piecewise aggregate approximation (PAA) [7], adaptive piecewise constant approximation (APCA) [8], and so forth use l adjacent segments to represent the time series with length m ($m \gg l$). Furthermore, perceptually important points (PIP) [9, 10], critical point model (CMP) [11], and so on reduce the dimensions by preserving the salient points. Another common family of time series representation approaches transform time series into discrete symbols and perform string operations on time series, for example, symbolic aggregate approximation (SAX) [12], shape description alphabet (SDA) [13], and other symbols generated method based on clustering [14, 15]. Representing time series in the transformation is another large family, such as discrete Fourier transform (DFT) [4] and discrete wavelet transform

(DWT) [16] which transform the original time series into frequency domain. After transformation, only the first few or the best few coefficients are chosen to represent the original time series [3]. Many of the representation schemes are incorporated with different multidimensional spatial indexing techniques (e.g., k-d tree [17] and r-tree and its variants [18, 19]) to index sequences to improve the query efficiency during similarity searching. Given two time series S and Q and their representations PS and PQ, a similarity measure function D calculates the distance between the two time series, denoted by $D(PQ, PS)$ to describe the similarity/dissimilarity between Q and S, such as Euclidean distance (ED) [4] and the other Lp norms, dynamic time warping (DTW) [20, 21], longest common subsequence (LCSS) [22], the slope distance [23], and the pattern distance [24].

The multidimensional time series similarity searches study mainly two aspects, the overall matching and match-by-dimension. The overall matching treats the MTS as a whole because of the important correlations of the variables in MTS datasets. Many overall matching similarity measures are based on principal component analysis (PCA). The original time series are represented by the eigenvectors and the eigenvalues after transformation. The distance between the eigenvectors weighted by eigenvalues is used to describe the similarity/dissimilarity, for example, Eros [3], S_{PCA} [25], and $S^{\lambda}_{\mathrm{PCA}}$ [26]. Lee and Choi [27] combined PCA with the hidden Markov model (HMM) to propose two methods, PCA + HMM and PCA + HMM + SVM, to find similar MTS. With the principal components such as the input of several HMMs, the similarity is calculated by combining the likelihood of each HMM. Guan et al. [28] proposed a pattern matching method based on point distribution (PD) for multivariate time series. Local important points of a multivariate time series and their distribution are used to construct the pattern vector. The Euclidean distance between the pattern vectors is used to measure the similarity of original time series.

By contrast, match-by-dimension breaks MTS into multiple univariate time series to process separately and then aggregates them to generate the result. Li et al. [29] searched the similarity of each dimensional series and then synthesized the similarity of each series by the traditional BORDA voting method to obtain the overall similarity of the multivariate time series. Compared to the overall matching, match-by-dimension could take advantage of present univariate time series similarity analysis approaches.

In this paper, a new algorithm based on the weighted BORDA voting method for the MTS k nearest neighbor (kNN) searching is proposed. Each MTS dimension series is considered as a separate univariate time series. Firstly, similarity searching approach is used to search the similarity sequence for each dimension series; then the similar sequences of each dimensional series are synthesized on the weighted BORDA voting method to generate the multivariate similar sequences. Compared to the measure in [29], our proposed method considers the dimension importance and the similarity gap between the similar sequences and generates more accurate similar sequences.

In the next section, we briefly describe the BORDA voting method and some similarity measures widely used. Section 3 presents the proposed algorithm to search the kNN sequences. Datasets and experimental results are demonstrated in Section 4. Finally, we conclude the paper in Section 5.

2. Related Work

In this section, we will briefly discuss BORDA voting method, the method in [29], and the DTW, on which our proposed techniques are based. Notations section contains the notations used in this paper.

2.1. BORDA Voting Method. BORDA voting, a classical voting method in group decision theory, is proposed by Jena-Charles de BORDA [30]. Supposing k is the number of winners, c is the number of candidates; e electors express their preference from high to low in the sort of candidates. To every elector's vote, the candidate ranked first is provided e points (called voting score), the second candidate e-1 points, followed by analogy, and the last one is provided 1 point. The accumulated voting score of the candidate is BORDA score. The candidates, BORDA scores in the top k, are called BORDA winners.

2.2. Similarity Measure on Traditional BORDA Voting. Li et al. [29] proposed a multivariate similarity measure based on BORDA voting, denoted by S_{BORDA}; the measure is divided into two parts: the first one is the similarity mining of univariate time series and the second one is the integration of the results obtained in the first stage by BORDA voting. In the first stage, a certain similarity measure is used to query kNN sequences on univariate series of each dimension in the MTS. In the second stage, the scores of each univariate similar sequence are provided through the rule of BORDA voting. The most similar sequence scores i points, the second scores i-1, followed by a decreasing order, and the last is 1. The sequences with same time period or very close time period will be found in different univariate time series. According to the election rule, the sequences whose votes are less than the half of dimension are eliminated; then the BORDA voting of the rest of sequences is calculated. If a sequence of some certain time period appears in the results of p univariate sequences and its scores are $s_1, s_2, \ldots, s_p$, respectively, then the similarity score of this sequence is the sum of all the scores. In the end, the sequence with the highest score is the most similar to the query sequence.

2.3. Dynamic Time Warping Distance. Dynamic programming is the theoretical basis for dynamic time warping (DTW). DTW is a nonlinear planning technique combining time and distance measure, which was firstly introduced to time series mining areas by Berndt and Clifford [20] to measure the similarity of two univariate time series. According to the minimum cost of time warping path, the DTW distance supports time axis stretching but does not meet the requirement of triangle inequality.

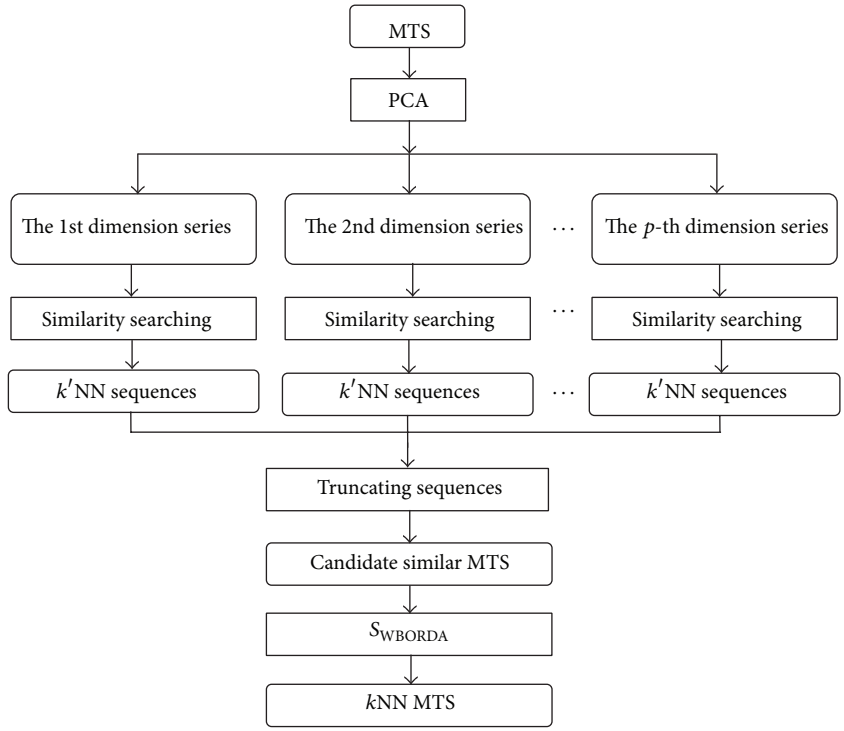

Figure 1: The model of similarity searching on weighted BORDA voting.

3. The Proposed Method

In the previous section, we have reviewed the similarity measure on traditional BORDA voting, S_{BORDA}, for multivariate time series. In this section, we propose a dimension-combination similarity measure based on weighted BORDA voting, called S_{WBORDA}, for MTS datasets kNN searching. The similarity measure can be applied for the whole sequence matching similarity searching and the subsequence matching similarity searching.

3.1. S_{WBORDA}: Multivariate Similarity Measure on Weighted BORDA Voting. S_{BORDA} takes just the order into consideration, without the actual similarity gap between two adjacent similar sequences that may lead to rank failure for the similar sequences. For example, assuming the four candidates r_1, r_2, r_3, r_4 take part in race, the first round position is r_1, r_2, r_3, r_4, the second is r_2, r_1, r_4, r_3, the third is r_4, r_3, r_1, r_2, and the last is r_3, r_4, r_2, r_1. The four runners are all ranked number 1 with traditional BORDA score (10 points), because of considering only the rank order, but without the speed gap of each runner in the race. In our proposed approach, we use the complete information of candidate, including the order and the actual gap to neighbor.

The multivariate data sequences S with n dimensions are divided into n univariate time series, and each dimension is a univariate time series. Given multivariate query sequence Q, to search the multivariate kNN sequences, each univariate time series is searched separately. For the jth dimension time series, the $k' + 1$ nearest neighbor sequences are $s_0, s_1, \ldots, s_{k'}$, where k' is equal or greater than k and s_0 is the jth dimension series of Q and is considered to be the most similar to itself. The distances between $s_1, \ldots, s_{k'}$ and s_0 are $d_1, \ldots, d_{k'}$, respectively, where d_{i-1} is less than or equal to d_i and $d_i - d_{i-1}$ describes the similarity gap between s_i and s_{i-1} to s_0. Let the weighted voting score of s_0 be $k' + 1$ and let $s_{k'}$ be 1; the weighted voting score of the sequence s_i, vs_i, is defined by

$$\text{vs}_i = w_j \left(1 + k' \left(1 - \frac{d_i}{d_{k'}}\right)\right) \quad \left(i = 1, \ldots, k' - 1\right), \tag{1}$$

where w is a weight vector based on the eigenvalues of the MTS dataset, $\sum_{j=1}^{n} w_j = 1$, and w_j represents the importance of the jth dimension series among the MTS. vs_i is inversely proportional to d_i; that is, s_0 is the baseline; the higher similarity gap between s_i and s_0 is, the lower weighted BORDA score s_i will get.

We accumulate the weighted voting score of each item in a candidate multivariate similar sequence and then obtain its weighted BORDA score. The candidate sequences are ranked on weighted BORDA scores, and the top k are the final similar sequences to Q. The model of similarity searching based on weighted BORDA voting is shown in Figure 1.

In the model of Figure 1, firstly, PCA is applied on original MTS and transforms it to new dataset Y whose variables are uncorrelated with each other. The first p dimensions series which contain most of characteristics of the original MTS are retained to reduce dimensions. Furthermore, univariate time series similarity searching is performed to each dimension

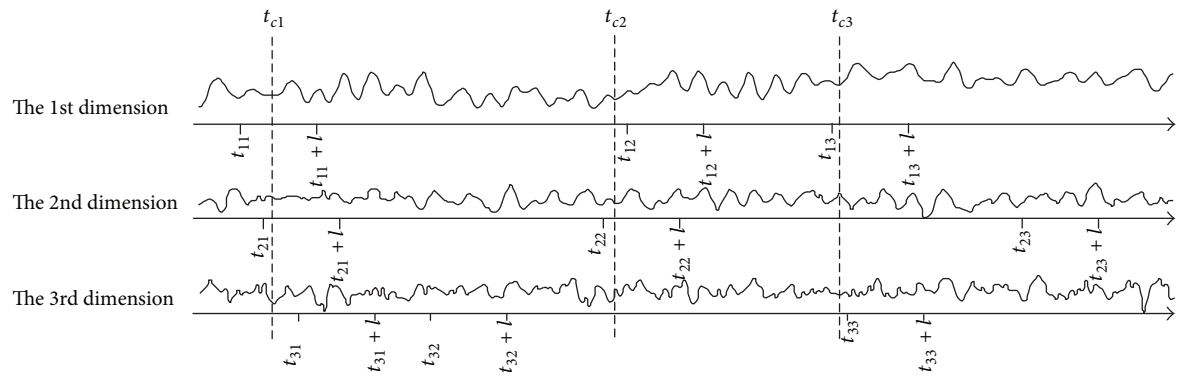

Figure 2: Truncating similar sequences for subsequence matching.

series in Y and finds out the univariate k'NN sequences; k' should be equal or greater than the final k. Moreover, k'NN sequences are truncated to obtain the candidate multivariate similar sequences. Finally, S_{WBORDA} is performed on candidate multivariate similar sequences to obtain the kNN of query sequences. Intuitively, S_{WBORDA} measures the similarity from different aspects (dimensions) and synthesizes them. The more aspects (dimensions) from measured sequences is similar to the query sequences, the more similar the sequence is to the query sequences of the whole. The following sections describe the similarity searching in detail.

3.2. Performing PCA on Original MTS. In our proposed method, all MTS dimension series are considered independent of each other, but, in fact, correlation exists among them more or less, so PCA is applied to the MTS which can be represented as a matrix $X_{m\times n}$ and m represents the length of series, and n is the number of dimensions (variables). Each row of X can be considered as a point in n-dimensional space. Intuitively, PCA transforms dataset X by rotating the original n-dimensional axes and generating a new set of axes. The principal components are the projected coordinate of X on the new axes [3].

Performing PCA on a multivariate dataset $X_{m\times n}$ is based on the correlation matrix or covariance matrix of X and results in two matrices, the eigenvectors matrix $C_{n\times n}$ and the variances matrix $L_{n\times 1}$. Each column of $C_{n\times n}$, called eigenvector, is a unit vector, geometrically, and it presents the new axes position in the original n-dimensional space. The variances matrix element $L_{i\times 1}$, called eigenvalue, provides the variance of the ith principal component. The matrix of the new projected coordinates $D_{m\times n}$ of the original data can be calculated by $D = X \cdot C$. The first dimension univariate time series of D is the first principal component and accounts for the largest part of variances presented in the original X; the ith principal component accounts for the largest part of the remaining variances and is orthogonal to the 1st, 2nd, ..., and $i-1$th dimensions. Select the first p components $D_{m\times p}$, which retain more than, for example, 90% of the total variation presented in the original data representing X. Thus, the dimensionality reduction may be achieved, as long as $p \ll n$. Geometrically, the original X is projected on the new p-dimensional space.

In whole sequence matching similarity searching, we apply PCA to all MTS items and retain p components so that more than, for example, 90% of the total variations are retained in all MTS items at least.

3.3. Truncating Univariate Similar Sequences. In candidate similar MTS, each dimension series starts at the same time. However, the similar sequences of each dimension time series may not start at the same time. The similar sequences with close start time of each dimension could be treated as in the same candidate similar MTS and truncated. The truncation includes four steps: grouping the sequences, deleting the isolated sequences, aligning the overlapping sequences, and reordering the sequences. After truncation, the candidate multivariate similar sequences could be obtained.

The truncation for whole sequence matching similarity searching is just a special case of subsequence matching, so we introduce the truncation for subsequence matching. In Figure 2, 3NN sequences are searched for multivariate query sequences with length l, and the application of PCA on the data MTS results in the principal component series with three dimensions. 3NN searching is performed on each dimension principal component series. The 3NN sequences of first dimension are s_{11} (the subsequence from t_{11} to $t_{11}+l$), s_{12} (from t_{12} to $t_{12}+l$), and s_{13} (from t_{13} to $t_{13}+l$). The univariate similar sequences are presented according to their occurrence time, and the present order does not reflect the similarity order to the query sequence. The 3NN sequences of the second dimension are s_{21} (from t_{21} to $t_{21}+l$), s_{22} (from t_{22} to $t_{22}+l$), and s_{23} (from t_{23} to $t_{23}+l$), and these of the third dimension are s_{31} (from t_{31} to $t_{31}+l$), s_{32} (from t_{32} to $t_{32}+l$), and s_{33} (from t_{33} to $t_{33}+l$).

(1) Grouping the Univariate Similar Sequences. The univariate similar sequences of all dimensions are divided into groups, so that in each group, for any sequence s, at least one sequence w, which overlaps with s over the half length of sequence l, could be found. The univariate similar sequence, which does

not overlap with any other similar sequences, will be put into a single group just including itself. In Figure 2, all the similar sequences are divided into five groups. The group g1 includes s_{11}, s_{21}, s_{31}. s_{11}, s_{21} overlaps with s_{21}, s_{31}, respectively, and the overlapping lengths are all over half of the length l, group g2 includes s_{32}, group g3 includes s_{12}, s_{22}, group g4 includes s_{13}, s_{33}, and group g5 includes s_{23}.

(2) Deleting the Isolated Sequences. The group, in which the number of similar sequences is less than half number of the dimensions, is called an isolated group, and the similar sequences in isolated group are called isolated similar sequences. In Figure 2, the number of similar sequences in group g2 or g5 is less than half of the number of dimensions, that is, 3, so the similar sequences in them are deleted.

(3) Aligning the Overlapping Sequences. The sequences in the same group are aligned to generate the candidate multivariate similar sequences. For one group, the average start time t of all the included sequences is calculated; then the subsequence from t to $t + l$, denoted by cs, is the candidate multivariate similar sequence. Each dimension series of cs is regarded as the univariate similar sequences. The similarity distance between cs and query sequence is recalculated by the selected univariate similarity measure dimension by dimension; if the group contains the ith dimension similar sequence, then the corresponding similarity distance is set to the ith dimension series of cs to reduce computation. In Figure 2, for group g1, the average of $t_{11}, t_{21}, t_{31} t_{c1}$ is calculated; then the subsequence s_{tc1}, from t_{c1} to $t_{c1} + l$, is the candidate multivariate similar sequence. For group g3, the similarity distance between the 2nd dimension series of s_{tc2} and the query sequence should be recalculated. The same alignment operation is performed on group g4 to obtain the candidate multivariate sequence s_{tc3}.

(4) Reordering the Candidate Similar Sequences. For each dimension, the candidate univariate similar sequences are reordered by the similarity distance calculated in Step (3). After reordering, S_{WBORDA} is used to synthesize the candidate similar sequences and generate the multivariate kNN sequences.

In whole matching kNN searching, the similar sequences are either whole overlapping or not overlapping each other at all, and the truncation steps are the same as those of the subsequence matching.

3.4. Computing Weights. By applying PCA to the MTS, the principal components series and the eigenvalues, which can represent the variances for principal components, are obtained. When we calculate the weighted BORDA score, we take into consideration both the similarity gap and the dimension importance for S_{WBORDA}. The heuristics proposed algorithm in [3] is used to calculate the weight vector w based on the eigenvalues. Variances are aggregated by a certain strategy, for example, min, mean, and max, on eigenvalues vectors dimension by dimension, and the vector $w\langle w_1, w_2, \ldots, w_k\rangle$ is obtained. The weight vector element is defined by

$$w_i = \frac{f(V_i)}{\sum_{j=1}^{k} f(V_j)} \qquad i = 1, \ldots, p, \tag{2}$$

where $f()$ denotes the aggregating strategy. V_i means the variance vector of ith dimension of MTS items. Generally, for subsequence matching, V_i includes one element, and whole matching is greater than 1. Intuitively, each w_i in the weight vector represents the aggregated variance for all the ith principal components. The original variance vector could be normalized before aggregation.

4. Experimental Evaluation

In order to evaluate the performance of our proposed techniques, we performed experiments on six real-world datasets. In this section, we first describe the datasets used in the experiments and the experiments methods followed by the results.

4.1. Datasets. The experiments have been conducted on four UCI datasets, [31] electroencephalogram (EEG), Australian sign language (AUSLAN), Japanese vowel (JV), and robot execution failure (REF), which are all labeled MTS datasets.

The EEG contains measurements from 64 electrodes placed on the scalp and sampled at 256 Hz to examine EEG correlates of genetic predisposition to alcoholism. Three versions, the small, the large, and the full, are included in this dataset according to the volume of the original data. We utilized the large dataset containing 600 samples and 2 classes.

The AUSLAN2 consists of samples of AUSLAN (Australian sign language) signs. 27 examples of each of 95 AUSLAN signs were captured from 22 sensors placed on the hands (gloves) of a native signer. In total, there are 2565 signs in the dataset.

The JV contains 640 time series of 12 LPC cepstrum coefficients taken from nine male speakers. The length of each time series is in the range 7–29. It describes the uttering of Japanese vowels /ae/ by a speaker successively. The dataset contains two parts: training and test data; we utilized the training data which contains 270 time series.

The REF contains force and torque measurements on a robot after failure detection. Each failure is characterized by 6 forces/torques and 15 force/torque samples. The dataset contains five subdatasets LP1, LP2, LP3, LP4, and LP5; each of them defines a different learning problem. The LP1, LP4, and LP5 subdatasets were utilized in the experiment. LP1 which defines the failures in approach to grasp position contains 88 instances and 4 classes, LP4 contains 117 instances and 3 classes, and LP5 contains 164 instances and 5 classes. A summary is shown in Table 1.

Table 1: Summary of datasets used in the experiments.

Dataset	Number of variables	Mean length	Number of instances	Number of classes
EEG	64	256	600	2
AUSLAN2	22	90	2565	95
JV	12	16	270	9
LP1	6	15	88	4
LP4	6	15	117	3
LP5	6	15	164	5

4.2. Method. In order to validate our proposed similarity measure S_{WBORDA}, 1NN classification, and 10-fold cross validation are performed [32]. That is, each dataset is divided into ten subsets, 1-fold for testing and the rest 9 for training. For each query item in the testing set, 1NN is searched in the training set and the query item is classified according to the label of the 1NNs, and the average precision is computed across all the testing items. The experiment is repeated 10 times for different testing set and training set, and 10 different error rates are obtained; then the average error across all 10 trials is computed for 10-fold cross validation. We performed 10 times 10-fold cross validation and computed the average error across all 10-fold cross validations to estimate the classification error rate. The similarity measures tested on our experiments include S_{BORDA}, PD, Eros, and S_{PCA}. DTW is selected as the univariate similarity measure for S_{BORDA} and S_{WBORDA}. They are denoted as $S_{\text{BORDA\_DTW}}$ and $S_{\text{WBORDA\_DTW}}$, respectively. For DTW, the maximum amount of warping Q is decreased to 5% of the length. DTW has been extensively employed in various applications and time series similarity searching, because DTW can be applied to two MTS items warped in the time axis and with different lengths.

All other measures except PD and Eros require determining the number of components p to be retained. Classification has been conducted for consecutive values of p which retain more than 90% of total variation, until the error rate reaches the minimum. The number which retains less than 90% of total variation is not considered. For Eros, the experiments are conducted as proposed in [3]. The Wilcoxon signed-rank test is used to ascertain if $S_{\text{BORDA\_DTW}}$ yields an improved classification performance on multiple data sets in general. PCA is performed on the covariance matrices of MTS datasets.

4.3. Results. The classification error rates are presented in Table 2 in the form of percentages. Although experiments have been conducted for various p, that is, the number of principal components (for $S_{\text{BORDA\_DTW}}$, $S_{\text{WBORDA\_DTW}}$, and S_{PCA}), only the best classification accuracies are presented.

Firstly, we will compare similarity measures with respect to each dataset, respectively. For the EEG dataset, S_{PCA} produces the best performance and performs significantly better than the others. With regard to the AUSLAN2 dataset, Eros produces the lowest classification error rate and PD gives very poor performance. For the overall matching method, for example, Eros, S_{PCA} performs better than the others. For the JV dataset, S_{PCA} gives the best performance and the $S_{\text{BORDA\_DTW}}$ makes the poorest performance. For the LP1 dataset, $S_{\text{WBORDA\_DTW}}$ makes the best performance. For the LP4 dataset, $S_{\text{WBORDA\_DTW}}$ makes the best performance and $S_{\text{BORDA\_DTW}}$ and $S_{\text{WBORDA\_DTW}}$ perform better than others. In the end, for the LP5 dataset, $S_{\text{WBORDA\_DTW}}$ gives the best performance.

Finally, the similarity measures are compared for all the datasets. Between $S_{\text{BORDA\_DTW}}$ and $S_{\text{WBORDA\_DTW}}$, the Wilcoxon signed-rank test reports that P value equals 0.043 (double side) and shows that the algorithms are significantly different. With 5% significance level, $S_{\text{WBORDA\_DTW}}$ has made better performance over $S_{\text{BORDA\_DTW}}$. Compared to Eros and S_{PCA}, the $S_{\text{WBORDA\_DTW}}$ has better performance on LP1, LP4, and LP5. But it shows poor performance on EEG, AULSAN2, and JV. Table 3 shows the number of principal components, which just retain more than 90% of total variation, in experiment datasets. For LP1, LP4, and LP5, the first few principal components retained most of the variation after PCA performing, but for EEG, AUSLAN2, and JV, to retain more than 90% of total variation, more principal components should be retained. $S_{\text{WBORDA\_DTW}}$ searches the similar sequences dimension by dimension and then synthesizes them; it is hard to generate the aligned candidate multivariate similar sequences when many principal components are contained in the principal component series. Furthermore, for the datasets, for example, EEG, AUSLAN2, and JV, the first few principal components could not retain sufficient information of original series, and $S_{\text{WBORDA\_DTW}}$ produces poor precision. $S_{\text{WBORDA\_DTW}}$ could make better performance on the datasets which aggregate the most variation in the first few principal components after PCA.

5. Conclusion and Future Work

A match-by-dimension similarity measure for MTS datasets, S_{WBORDA}, is proposed in this paper. This measure is based on principal component analysis, weighted BORDA voting method, and univariate time series similarity measure. In

Table 2: Classification error rate (%).

	EEG	AUSLAN2	JV	LP1	LP4	LP5
S_{BORDA_DTW}	$_{(14)}27.7$	$_{(4)}52.6$	$_{(5)}53.4$	$_{(3)}15.9$	$_{(3)}9.8$	$_{(4)}24.0$
S_{WBORDA_DTW}	$_{(14,max)}27.7$	$_{(4,max)}50.4$	$_{(6,max)}52.4$	$_{(4,mean)}13.0$	$_{(3,mean)}8.1$	$_{(3,mean)}21.3$
PD	38.4	68.9	45.2	14.1	12.3	41.2
Eros	$_{(max)}5.8$	$_{(mean)}11.1$	$_{(mean)}30.6$	$_{(max)}20.6$	$_{(mean)}10.3$	$_{(mean)}35.5$
S_{PCA}	$_{(14)}1.2$	$_{(4)}13.2$	$_{(12)}29.5$	$_{(3)}24.7$	$_{(4)}20.0$	$_{(3)}35.2$

(Numbers in parentheses indicate the p, i.e., the number of principal components retained, "max" and "mean" indicate the aggregating functions for weight w.)

Table 3: Contribution of principal components.

	EEG	AUSLAN2	JV	LP1	LP4	LP5
Number of variables	64	22	12	6	6	6
Number of retained principal components*	14	4	5	3	3	3
Retained variation (%)	90.1	92.7	93.9	92.4	94.8	91.2

(*For every dataset, if the number of principal components is less than the number in Table 3, the retained variation will be less than 90%.)

order to compute the similarity between two MTS items, S_{WBORDA} performs PCA on the MTS and retains k dimensions principal component series, which present more than 90% of the total variances. Then a univariate similar analysis is applied to each dimension series, and the univariate similar sequences are truncated to generate candidate multivariate similar sequences. At last, the candidate sequences are ordered by weighted BORDA score, and the kNN sequences are obtained. Experiments demonstrate that our proposed approach is suitable for small datasets. The experimental result has also shown that the proposed method is sensitive to the number of the classes in datasets. In the further work, it will be investigated furtherly.

In the literature, at present, there are still not so many studies in similarity analysis for MTS. In the future, we will explore new integration methods.

Notations

n: The number of MTS dimensions
m: The number of observations in a MTS
p: The number of retained principal components
k: The number of multivariate nearest neighbor sequences
k': The number of univariate nearest neighbor sequences, usually greater than or equal k
s_i: The ith univariate similar sequence, $1 \le i \le k'$
d_i: The distances between s_i and s_0, $1 \le i \le k'$
vs_i: Weighted voting score of the ith univariate similar sequence, $1 \le i \le k'$
$C_{n\times n}$: Eigenvectors matrix of PCA
$L_{n\times 1}$: Variances matrix of PCA
$f()$: The aggregating strategy of variances
w: Weight vector of principal components series
V_i: The ith dimension variances vector of MTS items, $1 \le i \le p$.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This research is supported by the Fundamental Research Funds for the Central Universities (no. 2009B22014) and the National Natural Science Foundation of China (no. 61170200, no. 61370091, and no. 61202097).

References

[1] X. L. Zhang, H. Begleiter, B. Porjesz, W. Wang, and A. Litke, "Event related potentials during object recognition tasks," *Brain Research Bulletin*, vol. 38, no. 6, pp. 531–538, 1995.

[2] M. W. Kadous, *Temporal Classification: Extending the Classification Paradigm to Multivariate Time Series*, The University of New South Wales, Kensington, Australia, 2002.

[3] K. Yang and C. Shahabi, "A PCA-based similarity measure for multivariate time series," in *Proceedings of the 2nd ACM International Workshop on Multimedia Databases (MMDB '04)*, pp. 65–74, November 2004.

[4] R. Agarwal, C. Faloutsos, and A. Swami, "Efficient similarity search in sequence databases," in *Proceedings of the International Conference on Foundations of Data Organization and Algorithms (FODO '93)*, pp. 69–84, 1993.

[5] E. Keogh and M. Pazzani, "An enhanced representation of time series which allows fast and accurate classification, clustering and relevance feedback," in *Proceedings of the 4th International Conference of Knowledge Discovery and Data Mining*, pp. 239–243, 1998.

[6] E. Keogh and P. Smyth, "A probabilistic approach to fast pattern matching in time series databases," in *Proceedings of the 3rd International Conference of Knowledge Discovery and Data Mining*, pp. 24–30, 1997.

[7] E. Keogh and M. Pazzani, "A Simple dimensionality reduction technique for fast similarity search in large time series databases," in *Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 122–133, 2000.

[8] E. Keogh, K. Chakrabarti, S. Mehrotra, and M. Pazzani, "Locally adaptive dimensionality reduction for indexing large time series databases," in *Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data*, vol. 30, 2, pp. 151–162, May 2001.

[9] F. L. Chung, T. C. Fu, R. Luk, and V. Ng, "Flexible time series pattern matching based on perceptually important points," in *Proceedings of the International Joint Conference on Artificial Intelligence Workshop on Learning from Temporal and Spatial Data*, pp. 1–7, 2001.

[10] T. Fu, "A review on time series data mining," *Engineering Applications of Artificial Intelligence*, vol. 24, no. 1, pp. 164–181, 2011.

[11] D. Bao, "A generalized model for financial time series representation and prediction," *Applied Intelligence*, vol. 29, no. 1, pp. 1–11, 2008.

[12] J. Lin, E. Keogh, S. Lonardi, and B. Chiu, "A symbolic representation of time series, with implications for streaming algorithms," in *Proceedings of the 8th ACM SIGMOD International Conference on Management of Data Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD '03)*, pp. 2–11, June 2003.

[13] H. A. Jonsson and Z. Badal, "Using signature files for querying time-series data," in *Proceedings of the 1st European Symposium on Principles and Practice of Knowledge Discovery in Databases*, pp. 211–220, 1997.

[14] G. Hebrail and B. Hugueney, "Symbolic representation of long time-series," in *Proceedings of the Applied Stochastic Models and Data Analysis Conference*, pp. 537–542, 2001.

[15] B. Hugueney and B. B. Meunier, "Time-series segmentation and symbolic representation, from process-monitoring to data-mining," in *Proceedings of the 7th International Conference on Computational Intelligence, Theory and Applications*, pp. 118–123, 2001.

[16] K. Chan and A. W. Fu, "Efficient time series matching by wavelets," in *Proceedings of the 1999 15th International Conference on Data Engineering (ICDE '99)*, pp. 126–133, March 1999.

[17] B. C. Ooi, K. J. McDonell, and R. Sacks-Davis, "Spatial kd-tree: an indexing mechanism for spatial databases," in *Proceedings of IEEE International Computers, Software, and Applications Conference (COMPSAC '87)*, pp. 433–438, 1987.

[18] A. Guttman, "R-trees: a dynamic index structure for spatial searching," in *Proceedings of the 1984 ACM SIGMOD international conference on Management of data (SIGMOD '84)*, 1984.

[19] N. Beckmann, H. Kriegel, R. Schneider, and B. Seeger, "R-tree. An efficient and robust access method for points and rectangles," in *Proceedings of the 1990 ACM SIGMOD International Conference on Management of Data*, pp. 322–331, May 1990.

[20] D. J. Berndt and J. Clifford, "Using dynamic time warping to find patterns in time series," in *Proceedings of the AAAI KDD Workshop*, vol. 10, pp. 359–370, 1994.

[21] E. Keogh and M. Pazzani, "Derivative dynamic time warping," in *Proceedings of the 1st SIAM International Conference on Data Mining*, 2001.

[22] M. Paterson and V. Dančík, *Longest Common Subsequences*, Springer, Berlin, Germany, 1994.

[23] J.-Y. Zhang, Q. Pan, P. Zhang, and J. Liang, "Similarity measuring method in time series based on slope," *Pattern Recognition and Artificial Intelligence*, vol. 20, no. 2, pp. 271–274, 2007.

[24] D. Wang and G. Rong, "Pattern distance of time series," *Journal of Zhejiang University*, vol. 38, no. 7, pp. 795–798, 2004.

[25] W. J. Krzanowski, "Between-groups comparison of principal components," *Journal of the American Statistical Association*, vol. 74, no. 367, pp. 703–707, 1979.

[26] M. C. Johannesmeyer, *Abnormal Situation Analysis Using Pattern Recognition Techniques and Historical Data*, University of California, Santa Barbara, Calif, USA, 1999.

[27] H. Lee and S. Choi, "PCA+HMM+SVM for EEG pattern classification," *Signal Processing and Its Applications*, vol. 1, no. 7, pp. 541–544, 2003.

[28] H. Guan, Q. Jiang, and S. Wang, "Pattern matching method based on point distribution for multivariate time series," *Journal of Software*, vol. 20, no. 1, pp. 67–79, 2009.

[29] S.-J. Li, Y.-L. Zhu, X.-H. Zhang, and D. Wan, "BORDA counting method based similarity analysis of multivariate hydrological time series," *Shuili Xuebao*, vol. 40, no. 3, pp. 378–384, 2009.

[30] D. Black, *The Theory of Committees and Elections*, Cambridge University Press, London, UK, 2nd edition, 1963.

[31] H. Begleiter, "UCI Machine Learning Repository," Irvine, CA: University of California, School of Information and Computer Science, 1999, http://mlr.cs.umass.edu/ml/datasets.html .

[32] F. Fábris, I. Drago, and F. M. Varejão, "A multi-measure nearest neighbor algorithm for time series classification," in *Proceedings of the 11th Ibero-American Conference on AI*, pp. 153–162, 2008.