Correlating Heterogeneous Time Series using Change Information

Chen Luo
Department of Computer Science
Rice University
Houston, TX, USA
cl67@rice.edu

Anshumali Shrivastava
Department of Computer Science
Rice University
Houston, TX, USA
anshumali@rice.edu

ABSTRACT

Calculating the correlation between different time series is an important data mining task. In recent years, time series data have two major properties: (1) Heterogeneity property (Different Time series may have different patterns.) and (2) High Dimensional and Large Scale. In addition, for heterogeneous time series with different pattern, they may also have high correlation. Despite their importance, there has been little previous work addressing the correlation between two types of heterogeneous time series data.

In this paper, we propose an approach that is capable of (1) evaluating evaluate the correlation between heterogeneous time series (time series with different patterns), and (2) dealing with very large scale problems. We investigated a change based correlation Coefficient to evaluate the correlation between different heterogeneous time series. In addition, we proposed hashing based method to do fast searching and clustering on very large scale time series. The experimental results on Synthetic data sets and real world data set show the effectiveness and efficiency of our algorithms.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Data Mining

General Terms

Application

Keywords

Correlation, Time Series, Hashing Learning

1. INTRODUCTION

Time series correlation is a major research topic in data mining area. Such correlating techniques has been applied in many real world problems. For example, some researchers use time series correlation techniques to analysis the signal information for speech processing [22]. Image processing researchers also use time series correlation techniques to deal with the image retrieval problems and object detection problems [29, 24]. In the system diagnose area[13, 25], time series correlation techniques also widely used to mine the system behavior and diagnose system failures. Time series correlation techniques can also be used for analyzing bio-sequences (e.g DNA Sequence, etc [17] etc.

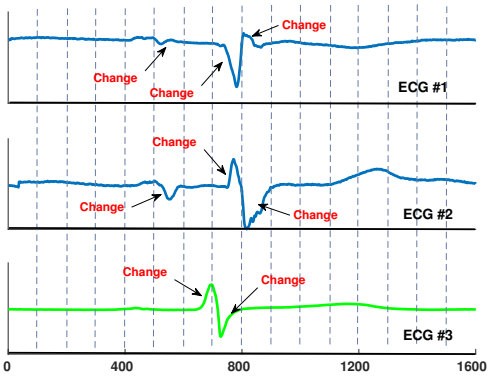

Figure 1: Three ECG time series with two labels

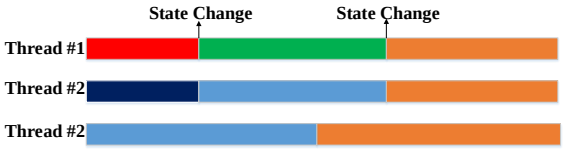

Figure 2: Three Thread time series

However, in most real world problems, time series data often have different patterns (Heterogeneous time series). For example, in the area of system analysis area. Each performance counter can be regarded as a time series (e.g. CPU Usage, Memory Usage, etc.). Some of the time series may be a periodical time series, but others may be a linear or random patterns. However, heterogeneous time series may also be correlations between each other. So, how to calculate the correlation between heterogeneous time series data is a challenge.

Analysis of ECG (Electrocardiogram) data

Electrocardiography [8] (ECG or EKG*) is the process of recording the electrical activity of the heart over a period of time using electrodes placed on a patient's body. These

SIGKDD '16 San Francisco, California USA

electrodes detect the tiny electrical changes on the skin that arise from the heart muscle depolarizing during each heartbeat. By analyzing such data, one can find some useful information hidden behind the human body, thus to uncover some miracle of human body [27].

Such ECG data can be regarded as time series data. Detect the correlation between each ECG time series is a powerful tool to analyze ECG data. After knowing the correlation between different ECG data, we can use such correlation to discover the hidden relation ship between each human and disease [14].

However, we can see from Fig.1 that, ECG time series data correlation are regarded as tiny electrical change at the same time [27]. And the change has different patterns. As a result, some classical similarity or correlation method can not deal with such problem well, even DTW some times also fail to detect such patterns. As a result, a change based correlation is needed.

Thread Behavior Mining in High Performance Computer. High performance computers (HPC) have become enormously complex. Today, the largest systems consist of more than tens of thousands of nodes. Nodes themselves are equipped with one or more multicore microprocessors[1].

As a result, it is increasingly difficult for application developers writing complex scientific programs to attain a significant fraction of peak performance on modern microprocessor-based computer systems. So, how to automatically analysis and monitoring the HPC is a major task for HPC researchers [15, 26].

HPCToolkit[1], introduced by Dr.John Mellor-Crummey, can generate some performance information of each process (or thread if application is multithreaded.) along the time. So, each process (thread) can be represented as a time series. (Depend on which aspect of a thread to be represent, domain knowledge required) The thread change information (e.g. change from one state to another state) can directly reflect some important properties of different threads.

Fig. 2 shows a example of three thread time series data. From the figure, we can see that above two thread often change at same time, so they may have highly correlation with each other. However, using the point to point based similarity method, we can not handle such heterogeneity property of different thread time series.

As showed in above examples, most of the existing time series similarity measures (e.g. L1-Distance, L2-Distance [7], and DTW-Distance [18], etc) or correlation measures (e.g. Pearson Correlation [19], Kendall rank correlation [11], and Spearman's rank correlation [20], etc.) can not deal with such heterogeneous properties of the time series. Because of that, for heterogeneous time series correlating, the correlation information is often associated with the change of time series during a time period rather than a point-to-point corresponding relationship in the traditional correlation analysis techniques. And the existing similarity and correlation measures only consider the point to point similarity or correlation between different time series. As a result, In order to deal with heterogeneity properties of time series with different patterns. We proposed a change based correlation coefficient. The intuition of this correlation is:

If two time series often change at the same time, they may have correlation with each other.

[1]http://hpctoolkit.org/

The detailed definition will be introduced in section 2. Our change based correlation method firstly extract the change information of the time series data, and then use the change information to calculate the correlation coefficient between the two time series. We also use the hashing methods speed up the correlation calculation between time series, as well as searching and clustering. So the computational cost of our coefficient is low, thus our method can also deal with the large scale property of the time series data.

The contribution of this paper is listed as follow:

1. Motivated by real applications, we investigate the correlation problem as between heterogeneous time series (Time Series with different patterns). To the best of our knowledge, this is the first attempt to evaluate the correlation between time series with different patterns.

2. We proposed a correlation coefficient between heterogeneous time series, and use hashing method to speed up the calculating of the coefficient between time series, as well as the clustering and nearest neighbors search tasks.

3. The experiments on Synthetic data show the effectiveness and efficiency of our method.

The rest of the paper is organized as follows: In Section 2, we introduce the problem statement and formulation. Our approach is proposed in Section 3. The Empirical evaluation is shown in Section 4. In Section 5, we introduce some related works. Finally, we conclude our work in Section 6.

2. BACKGROUND

2.1 Preliminary Definition

In this section, we formally define some concept of this work, including Time Series, Change Point, Change Point set, Time Series Correlation.

DEFINITION 1 (TIME SERIES). *A time series, denoted as $S = (s_1, s_2, ..., s_m)$, where m is the number of points in the time series. The timestamps of a time series, denoted as $TS = (t(s1), t(s2), ..., t(sn))$, have the relationship of $t(s_i) = t(s_{i-1}) + \tau$, where $\tau$is the sampling interval.*

In this work, we consider the change information of a time series, so the change point of a time series is defined as follow:

DEFINITION 2 (CHANGE POINT). *A time series, denoted as $S = (s_1, s_2, ..., s_m)$, a change point is a time stamp $t_s(i)$that there is a change before and after this time stamp. Change contains the following types: mean change, variance change, frequency change, and the combination between them.*

The definition of change point set is defined as follow:

DEFINITION 3 (CHANGE POINT SET). *A time series, denoted as $S = (s_1, s_2, ..., s_m)$ The timestamps of a time series, denoted as $TS = (t(s1), t(s2), ..., t(sn))$ The change points set is denoted as $C_X = (t_x(1), t_x(2), ..., t_x(p))$ Where $t_x(i)$ denotes the change points of time series S.*

2.2 Change based Time Series Correlation

After define the time series, we define the correlation of this work as follow:

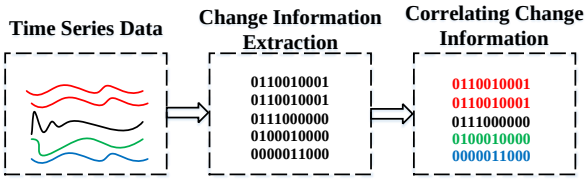

Figure 3: Overview of the Framework

Definition 4 (Change Based Correlation). *Suppose we have two time series:* $X_1 = (x_1, x_2, ..., x_m), Y_1 = (y_1, y_2, ..., y_m)$, *The change point set of* X *and* Y *are denoted as:* $C_X = (t_x(1), t_x(2), ..., t_x(p))$ $C_Y = (t_y(1), t_y(2), ..., t_y(q))$ *where* q *and* p *are numbers of change points for time series* X *and* Y. *Then if* X *and* Y *are correlated if and only if:*

$$\begin{cases} p = q \\ L1(C_X, C_Y) < \xi \end{cases} \tag{1}$$

where $L1(C_X, C_Y)$ *denotes the L1 distance, and* ξ *is the threshold of Time series correlation.*

3. THE APPROACH

In this section, we first propose a framework to analyze the correlation of heterogeneous time series, and then we introduce how to use hashing method to do fast searching.

3.1 Change Based Correlation Coefficient

As we introduced in Section. 2, change information is important for calculating the change based correlation method.

So, Given a time series, the first thing need to do is to extract the change information of the time series. In this work, we regard the change information as bit-stream. In the bit-stream, 1 denotes there is a change in this sub-series, and 0 if not.

After obtaining the change information of the time series, we then calculate the Jaccard similarity[7] coefficient between each other. // Here need to say something about why.

The Jaccard similarity between each bit-stream will be the correlation coefficient between these two time series. The framework of this correlation is showed in Fig.3.

3.2 Change Information Extraction

As we introduced in Section.2, change based correlation corresponds to the change information of the time series. Change information of a time series is a time period information, not a time point information. As a result, in order to extract the change information of the time series, we need to find the information in small time period of the time series (A sub-series).

Given a time series $S = (s_1, s_2, ..., s_m)$, where m is the number of points in the time series. Given a subs-series length k. The change information of the time series S can be represented as a bit-stream:

$B_S = \{b_0, b_1, ..., b_n\}$, where each b_i corresponds to a sub-series of length k for the original time series S as showed in Fig.??.

Given a sub-series $l^j = \{s_i, s_{i+1}, s_{i+2}, ..., s_{i+k-1}\}$, where w is the length of the sub-series. Then the change information of sub-series l is denoted as follow:

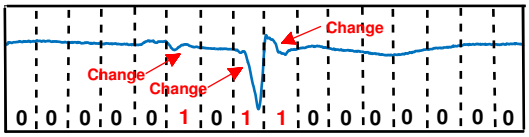

Figure 4: Change Information Extraction

$$b^l = \begin{cases} 1 & Have\ change\ in\ l \\ 0 & No\ change\ in\ l \end{cases} \tag{2}$$

As showed in Equ.2, the change information is the information that whether there's a change in the sub-series. In order to denote whether there's a change in the sub-series, we need to know to to detect change in the sub-series. Fig.4 shows how to extract the change information of the time series.

3.3 Change Detection

So,the problem here is: Given a sub-series $l^j = \{s_i, s_{i+1}, s_{i+2}, ..., s_{i+k-1}\}$,

how to denote whether there is a change or not in this time series. There are a so many time series change detection methods [12, 2] proposed in the literature.

In this work, the change detection task here is not find the change points of the time series, instead we only need to denote whether there's a change in the time series. It is pointed out that all the change point detection methods can be used here to detect change information. In our experiment, we use the the following method to detect change:

We equally divide the time series into two series:
$l^j_{Front} = \{s_i, s_{i+1}, s_{i+2}, ..., s_{i+(k-1)/2-1}\}$
and,
$l^j_{Front} = \{s_{i+(k-1)/2-1}, s_{i+(k-1)/2}, ..., s_{i+k-1}\}$.

So, regard l^j_{Front} and l^j_{Rear} as two data sampled from two distributions P_1 and P_2. So, if P_1 and P_2 are statistically the same, then we can say there's not change between each other. Otherwise, there is a change in this dataset.

Then, the problem here becomes a *Two Sample Problem* [5]. We use the Two Sample t-test [16] method to solve this problem:

Here, the t_{score} between l^j_{Front} and l^j_{Rear} can be calculated as:

$$t_{score} = \frac{\overline{l^j_{Front}} - \overline{l^j_{Rear}}}{\sigma_p \sqrt{2/k}} \tag{3}$$

where, $\overline{l^j_{Front}}$ and $\overline{l^j_{Front}}$ are the mean values of l^j_{Front} and l^j_{Front}. And σ_p is as follow:

$$\sigma_p = \frac{(k-1)\sigma^2_{l^j_{Front}} + (k-1)\sigma^2_{l^j_{Rear}}}{k-1} \tag{4}$$

Then, if $t_{score} > \alpha$, we can say that these two samples are from different distributions, and thus there is a change in the sub-series l^j.

3.4 Jaccard Similarity Coefficient

After obtaining the change information (Bit-stream) of each data, we then use Jacord Similarity Coefficient to calculate the Change Correlation of each time series.

The Jaccard Similarity [7] is defined as follow: Given two Bit-stream X and Y, the Jaccord distance is showed as follow:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{5}$$

where, $|A \cap B|$ denotes the number of bit that X and Y both 1. And $|A \cup B|$ denotes the number of bit that at least X or Y is 1. For example, given two bit stream: $X = 100111$, and $Y = 001110$. Then $|A \cap B| = 2$, and $|A \cap B| = 5$, so $J(A,B) = \frac{2}{5} = 0.4$.

3.5 Speed up Top-k Searching using LSH

As we introduced before, in the real world problem, the scale of time series data often huge. Mining such huge number of time series is a big challenge.

We can see from the change information that, it is a bit-stream data, so, we can use hashing method to speed up the searching method under change based correlation coefficient.

In this work, we LSH (Locality Sensitive Hashing) [9] to do fast search. // Exaplain the Fast search algorithm here.

3.6 Discuss about Sub-series Length w

In this research, the sub-series length w is a very important parameter. If the sub-series length w is too short, then the change information can not be captured. On the other hand, if the sub-series length w is too long, then there will be to much noise information.

In some cases, the value of k can be selected based on domain knowledge and experiments. However, in most real world situations, there are millions of time series and events, and we do not have enough domain knowledge to pre-select the values of all sub-series lengths. In this research, we use the parameter as our previous research work on Correlating Event with time series [13]. This method can auto-select the sub-series length for a time series based on the autocorrelation function [6] of the time series. Given a time series $S = (s_1, s_2, ..., s_n)$, the autocorrelation is showed as follow:

$$R(l) = E(s_i * s_{i-l}). \tag{6}$$

where l denotes the lag of the correlation. The autocorrelation function of a time series can be used to represent the energy of signals in the time series with a period of l [6]. Therefore, our length k can be assigned as the value of the first peak to include the significant signal of the time series.

4. EMPIRICAL EVALUATION

In this section, we make an empirical evaluation of our algorithm by performing a set of experiments on the synthetic data set.

4.1 Comparison Methods

In order to evaluate the effectiveness of our algorithm, we choose three time series similarity algorithms and four correlation coefficient in our experiment.

For the three similarity algorithm, we choose L1-Distance, L2-Distance [7], and DTW-Distance [18]. And for the three similarity algorithm, we choose Pearson correlation [4], which is the widely used methods for correlation mining in time series. In the rest of this subsection, we brief introduce the three similarity measures and the three Correlation measures.

4.1.1 Similarity Measures

In this work, we introduce three similarity measures between time series. Given a two time series $X = (x_1, x_2, ..., x_m), Y = (y_1, y_2, ..., y_m)$.

The L1-distance is showed as follow:

$$L1(X,Y) = \sum_{1}^{m} |x_i - y_i| \tag{7}$$

The L2-distance is showed as follow:

$$L1(X,Y) = \sqrt{\sum_{1}^{m} |x_i - y_i|^2} \tag{8}$$

The DTW distance is a famous time series similarity measure.

In order to introduce the DTW distance, we first construct an $m - by - m$ matrix W, where the $(i - th, j - th)$ element of the matrix W. The DTW distance is to find a path through the matrix that minimizes the total cumulative distance between X and Y. So, the optimal path is the path that minimize the warping cose:

$$DTW(X,Y) = \min \sqrt{\sum_{k=1}^{K} w_k} \tag{9}$$

where, w_k belongs to the $k - th$ element of a warping path P, which is a contiguous set of elements that represent a mapping between X and Y.

4.1.2 Correlation Measures

In this subsection, we introduce three widely used correlation measures between time series: Pearson Correlation [19], Kendall rank correlation [11], and Spearman's rank correlation [20].

The Pearson correlation method is one of the most widely used method for measuring the correlation between two time series. The Pearson correlation coefficient, denoted as $\rho.$ is calculated as follow:

$$\rho_{X,Y}^{Pearson} = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

where cov is the covariance, σ_X is the standard deviation of X, μ_X is the mean of X and $E[*]$ denotes the expectation.

The Kendall rank correlation [11] is defined as follow:

$$\rho_{X,Y}^{Kendall} = \frac{N_c - N_d}{m(m-1)/2}$$

Where N_c is the number of concordant pairs, and N_d is the number of discordant pairs, and m is the dimension of the time series. For any pair (x_i, y_i) and x_j, y_j, where $i \neq j$, are said to be concordant both $x_i > x_j$ and $y_i > y_j$, or $x_i < x_j$ and $y_i < y_j$. Otherwise, they are discordant.

The Spearman's rank correlation [20] is defined as follow:

$$\rho_{X,Y}^{Spearman} = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

where d_i is defined as the difference between the ranks of x_i and y_i.

Table 1: Summery of Synthetic Data

Change Type
Mean Change
Variance Change
Frequency Change + Variance Change
Mean Change + Frequency Change
Frequency Change + Variance Change
Mean Change + Frequency Change + Variance Change

4.2 Effectiveness Study on Synthetic Dataset

In this section, we introduce the experiment on the synthetic Dataset.

4.2.1 The Synthetic Dataset

Synthetic data set is very useful for evaluating algorithms and functions for data mining models[7]. In this section, we introduce the Synthetic Dataset used in our experiment.

In this synthetic Dataset, there are two patterns of time series: (1) Periodical Pattern, (2) Linear Pattern. Then, after obtaining the different pattern time series, we add white noise for each time series. Seven different types of changes are added randomly into each time series, the seven change types are showed in Table.1.

We generate a large data set, with 5 clusters, and the time series in the same cluster often change at the same time (The change point in the same cluster is added at the same time). Then, from this large data set, in order to test both the efficient and effective of our coefficient compared with other coefficient, we make the data set size from small to large, and also the time series length from short to long. Thus we extract 9 small data set, the scale of the dataset is showed in Table.3.

4.2.2 Clustering Task

In order to evaluate the performance of our correlation coefficient, we design a clustering task. In this experiment, we only use the Hierarchical Clustering [7] to evaluate the clustering performance.

Two evaluation methods are used for testing the clustering result: Accuracy [7], which is calculated as the percentage of target objects clustered into the correct clusters; and Normalized Mutual Information (NMI) [7], which is one of the most popular evaluation methods to evaluate the quality of clustering results.

From Table.2, we can see that, for the change correlation coefficient, the clustering result performance is better for the high dimensional data set. This is because that for high dimensional time series data, the proposed coefficient can extract more change information from the time series data and also make more accurate evaluation of the correlation. From this point of view, change based correlation is more suitable for the high dimensional time series data set. On the other hand, Change based correlation can obtain more accuracy results in different dataset compared with both the similarity method and the correlation coefficient. So, the result of clustering show the effectiveness of our coefficient.

Fig.5 shows the Execution time of the clustering task on each data set. From the result, we can see that change based correlation performed much faster than other algorithms. This is because, we only calculate the correlation on the extracted change information (Bit-stream), the calculating of change based correlation measure can be much faster than other similarity and correlation methods.

4.2.3 K-nearest Neighbors Task

We compute precision and recall [21] on the data-set using the LSH method, and other methods using the naive K-nearest Neighbors method. If the Searched time series is in the same cluster of the query time series, we regard it as a relevant items, and vice verse. We range the K from 1 to the cluster size.

For each top-k Nearest Neighbors search, the precision can be calculated as follow:

$$Precision = \frac{relevant\ item}{K} \tag{10}$$

and the recall can be calculated as follow:

$$Precision = \frac{relevant\ item}{Relevant\ Cluster\ Size} \tag{11}$$

The plots for all the three datasets are shown in Figure.6. We can clearly see that our proposed Change-based Correlation method gives significantly higher precision recall curves than other similarity and correlation methods. In addition the results are consistent across datasets.

Fig.7 shows the execution time by vary the data size and the time series length. In left one of Fig.7, we fix the value of time series length, and vary the data size n. We can see that the CPU execution time of other similarity methods increased sharply by enlarging the data size. And, the change based correlation do not change so much by enlarge the data size.

In right of Fig. 7, we fix the size of data size, and vary the value of time series length. Based on the results, we can see that the running time of the proposed change based method with LSH doesn't so much with the increase of time series length, while other methods increase by enlarging the time series length.

4.3 Effectiveness Study on Real Datasets

In this section, we will compare the proposed algorithm with the baseline algorithms on two real data sets.

4.3.1 Electrocardiogram Data set

The first real world dataset is ECG (Electrocardiogram) time series data set. This data set is comming from the the UCR time series Data set Archive [3].

Electrocardiography [8] (ECG or EKG*) is the process of recording the electrical activity of the heart over a period of time using electrodes placed on a patient's body. These electrodes detect the tiny electrical changes on the skin that arise from the heart muscle depolarizing during each heartbeat. By analyzing such data, one can find some useful information hidden behind the human body, thus to uncover some miracle of human body [27].

Such ECG data can be regarded as time series data. Detect the correlation between each ECG time series is a powerful tool to analyze ECG data. After knowing the correlation between different ECG data, we can use such correlation to discover the hidden relation ship between each human and disease [14].

However, we can see from Fig.1 that, ECG time series data correlation are regarded as tiny electrical change at

Table 2: Clustering Performance on Synthetic Data Set

Dataset	Measure	Proposed	$L1$	$L2$	DTW	Pearson	Kendall	Spearman
Sythetic-T0	Accuracy	**.854 ± .032**	.241 ± .098	.281 ± .012	.230 ± .061	.309 ± .140	.353 ± .026	.297 ± .036
	NMI	**.808 ± .034**	.026 ± .067	.076 ± .023	.028 ± .075	.140 ± .55	.395 ± .015	.150 ± .088
Sythetic-T1	Accuracy	**.838 ± .025**	.247 ± .026	.262 ± .032	.283 ± .012	.240 ± .018	.374 ± .067	.341 ± .067
	NMI	**.701 ± .030**	.003 ± .062	.057 ± .043	.064 ± .036	.046 ± .084	.404 ± .023	.230 ± .042
Sythetic-T2	Accuracy	**.806 ± .029**	.254 ± .066	.263 ± .080	.304 ± .022	.388 ± .024	.384 ± .032	.502 ± .182
	NMI	**.889 ± .012**	.028 ± .042	.056 ± .056	.054 ± .032	.303 ± .064	.394 ± .052	.450 ± .049
Sythetic-T3	Accuracy	**.856 ± .077**	.225 ± .028	.229 ± .034	.284 ± .062	.454 ± .032	.454 ± .032	.454 ± .032
	NMI	**.891 ± .017**	.021 ± .040	.041 ± .043	.086 ± .038	.454 ± .032	.454 ± .032	.454 ± .032

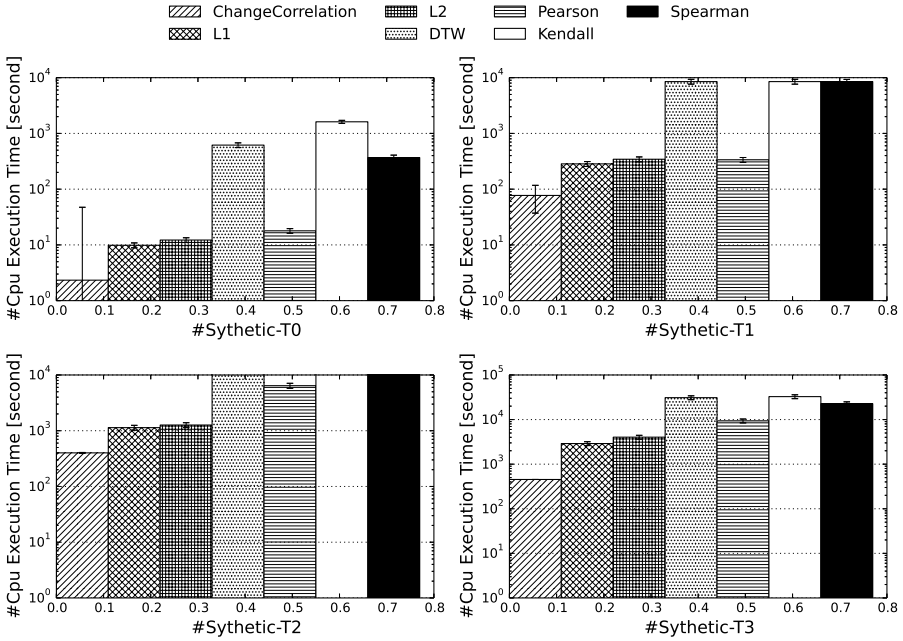

Figure 5: Top-k Nearest Neighbor Search

Table 3: Summery of Synthetic Data

DataSet	Data Size	Time Series Length
Synthetic-T0	1000	800
Synthetic-T1	1000	5000
Synthetic-T2	10000	800
Synthetic-T3	10000	5000

Table 4: Summary of the Four ECG Data Set

Data Set	Data Size	Time Series Length
CinC_ECG_torso	1380	1639
ECGFiveDays	861	136
TwoLeadECG	1139	82
ECG5000	4500	140

the same time [27]. And the change has different patterns. As a result, some classical similarity or correlation method can not deal with such problem well, even DTW some times also fail to detect such patterns. As a result, a change based correlation is needed.

4.3.2 HPC Thread Time Series Data set

The second real world dataset is from the HPC-tool Kit Dataset.

High performance computers (HPC) have become enormously complex. Today, the largest systems consist of more than tens of thousands of nodes. Nodes themselves are equipped with one or more multicore microprocessors[1].

As a result, it is increasingly difficult for application developers writing complex scientific programs to attain a significant fraction of peak performance on modern microprocessor-based computer systems. So, how to automatically analysis and monitoring the HPC is a major task for HPC researchers [15, 26].

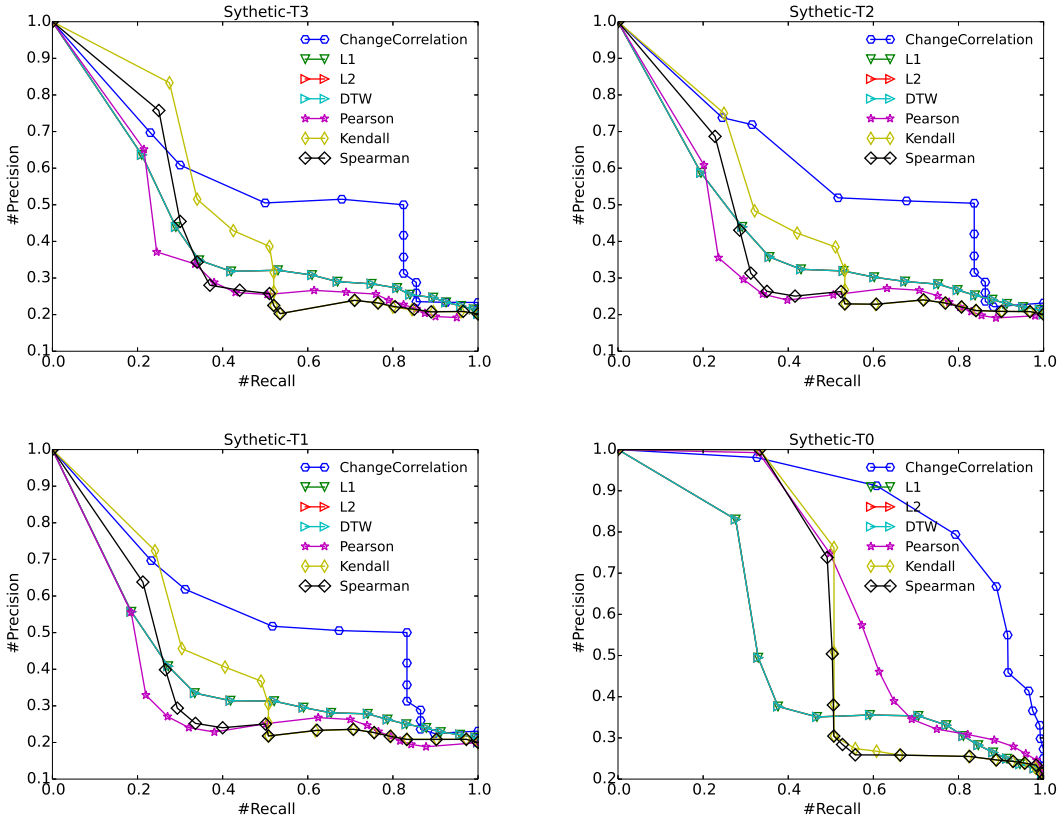

Figure 6: Precision Recall Curve for Different Algorithms

HPCToolkit[2], introduced by Dr.John Mellor-Crummey, can generate some performance information of each process (or thread if application is multithreaded.) along the time. So, each process (thread) can be represented as a time series. (Depend on which aspect of a thread to be represent, domain knowledge required) The thread change information (e.g. change from one state to another state) can directly reflect some important properties of different threads.

Fig. 2 shows a example of three thread time series data. From the figure, we can see that above two thread often change at same time, so they may have highly correlation with each other. However, using the point to point based similarity method, we can not handle such heterogeneity property of different thread time series.

5. RELATED WORK

In this section, we brief introduce some related works of our research.

5.1 Correlation between Time Series Data

Correlation between two time series has been widely studied, and some of them have been included in text books [10]. Pearson Correlation [4] is a basic correlation measure between time series, which has been widely used in practice [30]. Some extensions of Pearson correlation are also widely used. For example, lagged correlation is an extension to correlate a lagged dataset with another unlagged dataset using the Pearson product-moment method. In [28], the author uses the lagged-correlation to estimate the lead relationship between a set of time series. Because Pearson correlation is sensitive outliers in data set, Spearman Rank correlation and Kendall Rank correlation have been used in some scenarios [23] to overcome the drawbacks of Pearson correlation. In Spearman correlation, data is first sorted and each value assigned a rank, e.g., 1 is assigned to the lowest value. Spearman Rank correlation is calculated by taking the Pearson product-moment correlation of the ranks of the datasets. Kendall correlation is used to measure the similarity of the orderings of the data when ranked by each of data values. Because there is no ordering relationship among the different events, the above rank based algorithms cannot be directly used in our scenario.

[2]http://hpctoolkit.org/

5.2 LSH Hashing

Correlation between two time series has been widely studied, and some of them have been included in text books [10]. Pearson Correlation [4] is a basic correlation measure between time series, which has been widely used in practice [30]. Some extensions of Pearson correlation are also widely used. For example, lagged correlation is an extension to correlate a lagged dataset with another unlagged dataset using the Pearson product-moment method. In [28], the author uses the lagged-correlation to estimate the lead relationship between a set of time series. Because Pearson correlation is sensitive outliers in data set, Spearman Rank correlation and Kendall Rank correlation have been used in some sce-

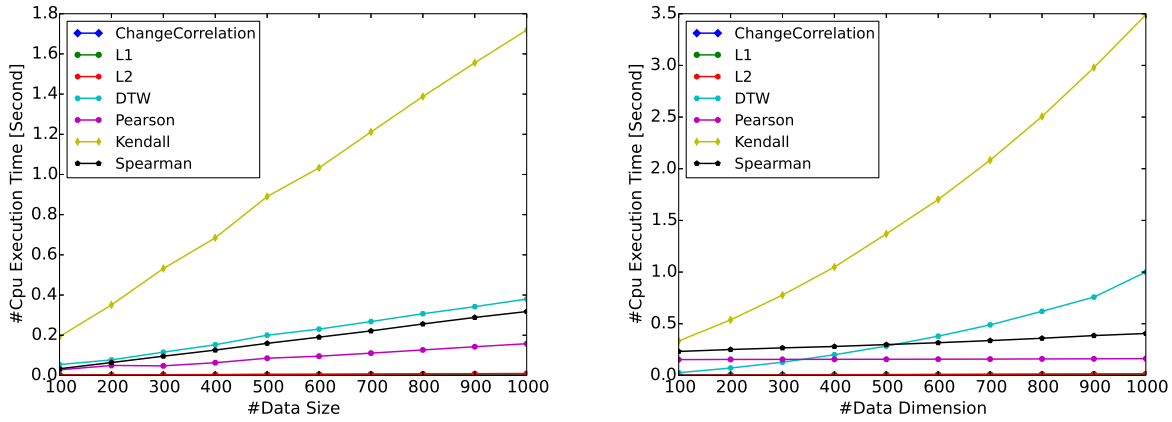

Figure 7: Efficiency by varying data size and dimension size

Table 5: Clustering Performance on Synthetic ECG Data Set From UCR Time Series Archive

Dataset	Measure	Proposed	$L1$	$L2$	DTW	Pearson	Kendall	Spearman
CinC_ECG_torso	Accuracy	**.839 ± .011**	.667 ± .068	.557 ± .012	.610 ± .061	.531 ± .140	.507 ± .019	.504 ± .013
	NMI	**.489 ± .019**	.236 ± .035	.019 ± .023	.010 ± .075	.280 ± .55	.150 ± .015	.049 ± .012
EGG_5000	Accuracy	**.538 ± .025**	.247 ± .026	.262 ± .032	.283 ± .012	.240 ± .018	.374 ± .067	.341 ± .067
	NMI	**.401 ± .030**	.003 ± .062	.057 ± .043	.064 ± .036	.046 ± .084	.404 ± .023	.230 ± .042
TwoLeadECG	Accuracy	**.810 ± .029**	.504 ± .066	.538 ± .080	.620 ± .022	.528 ± .064	.531 ± .052	.519 ± .049
	NMI	**.680 ± .012**	.081 ± .042	.043 ± .056	.137 ± .032	.047 ± .064	.062 ± .052	.074 ± .049
ECGFiveDays	Accuracy	**.832 ± .077**	.502 ± .028	.527 ± .034	.615 ± .062	.506 ± .032	.547 ± .032	.519 ± .032
	NMI	**.765 ± .017**	.002 ± .040	.002 ± .043	.361 ± .038	.075 ± .032	.023 ± .032	.086 ± .032

Table 6: Summary of the HPC Time Series Data Set

Data Set	Data Size	Time Series Length
Single PC	24	4096
MADNESS	264	32768

narios [23] to overcome the drawbacks of Pearson correlation. In Spearman correlation, data is first sorted and each value assigned a rank, e.g., 1 is assigned to the lowest value. Spearman Rank correlation is calculated by taking the Pearson product-moment correlation of the ranks of the datasets. Kendall correlation is used to measure the similarity of the orderings of the data when ranked by each of data values. Because there is no ordering relationship among the different events, the above rank based algorithms cannot be directly used in our scenario.

6. CONCLUSION AND FUTURE WORKS

Calculating the correlation between different time series is an important data mining task. In recent years, time series data have two major properties: (1) Heterogeneity property (Different Time series may have different patterns.) and (2) High Dimensional and Large Scale. In addition, for heterogeneous time series with different pattern, they may also have high correlation. Despite their importance, there has been little previous work addressing the correlation between two types of heterogeneous time series data.

In this paper, we propose an approach that is capable of (1) evaluating evaluate the correlation between heterogeneous time series (time series with different patterns), and (2) dealing with very large scale problems. We investigated a change based correlation Coefficient to evaluate the correlation between different heterogeneous time series. In addition, we proposed hashing based method to do fast searching and clustering on very large scale time series. The experimental results on Synthetic data sets and real world data set show the effectiveness and efficiency of our algorithms.

7. REFERENCES

[1] L. Adhianto, S. Banerjee, M. Fagan, M. Krentel, G. Marin, J. Mellor-Crummey, and N. R. Tallent. Hpctoolkit: Tools for performance analysis of optimized parallel programs. *Concurrency and Computation: Practice and Experience*, 22(6):685–701, 2010.

[2] X. C. Chen, K. Steinhaeuser, S. Boriah, S. Chatterjee, and V. Kumar. Contextual time series change detection. In *SDM*, pages 503–511. SIAM, 2013.

[3] Y. Chen, E. Keogh, B. Hu, N. Begum, A. Bagnall, A. Mueen, and G. Batista. The ucr time series classification archive, July 2015. `www.cs.ucr.edu/~eamonn/time_series_data/`.

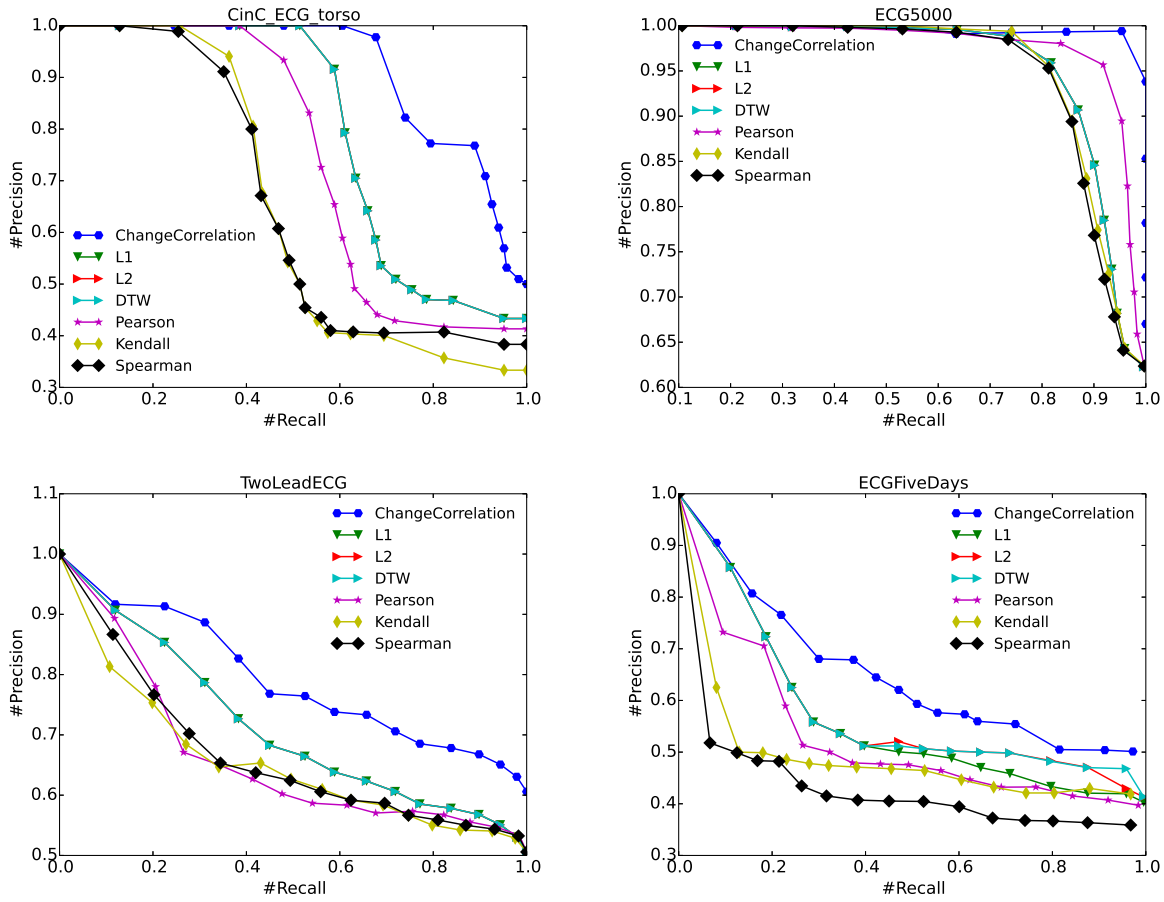

Figure 8: Precision Recall Curve (higher is better). We compare Change based correlation coefficient with other methods.

[4] J. Cohen. Statistical power analysis for the behavioral sciences. 1988.

[5] A. Gretton, K. M. Borgwardt, M. Rasch, B. Schölkopf, and A. J. Smola. A kernel method for the two-sample-problem. In *Advances in neural information processing systems*, pages 513–520, 2006.

[6] J. D. Hamilton. *Time series analysis*, volume 2. Princeton university press Princeton, 1994.

[7] J. Han, M. Kamber, and J. Pei. *Data mining: concepts and techniques: concepts and techniques*. Elsevier, 2011.

[8] N. J. Holter. New method for heart studies continuous electrocardiography of active subjects over long periods is now practical. *Science*, 134(3486):1214–1220, 1961.

[9] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

[10] R. A. Johnson and D. W. Wichern. *Applied multivariate statistical analysis*. Pearson, 2007.

[11] M. G. Kendall. A new measure of rank correlation. *Biometrika*, pages 81–93, 1938.

[12] S. Liu, M. Yamada, N. Collier, and M. Sugiyama. Change-point detection in time-series data by relative density-ratio estimation. *Neural Networks*, 43:72–83, 2013.

[13] C. Luo, J.-G. Lou, Q. Lin, Q. Fu, R. Ding, D. Zhang, and Z. Wang. Correlating events with time series for incident diagnosis. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1583–1592. ACM, 2014.

[14] H. J. L. Marriott and G. S. Wagner. *Practical electrocardiography*. Williams & Wilkins Baltimore, 1988.

[15] C. McCurdy and J. Vetter. Memphis: Finding and fixing numa-related performance problems on multi-core platforms. In *Performance Analysis of Systems & Software (ISPASS), 2010 IEEE International Symposium on*, pages 87–96. IEEE, 2010.

[16] D. S. Moore. *The basic practice of statistics*, volume 2. WH Freeman New York, 2007.

[17] D. W. Mount and D. W. Mount. *Bioinformatics: sequence and genome analysis*, volume 2. Cold spring harbor laboratory press New York:, 2001.

[18] M. Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[19] K. Pearson. *Mathematical contributions to the theory of evolution*, volume 13. Dulau and co., 1904.

Table 7: Clustering Performance on Synthetic ECG Data Set From UCR Time Series Archive

Dataset	Measure	Proposed	$L1$	$L2$	DTW	Pearson	Kendall	Spearman
Single PC	Accuracy	**.917 ± .011**	.835 ± .068	.815 ± .012	.870 ± .061	.501 ± .140	.527 ± .019	.514 ± .013
	NMI	**.889 ± .019**	.736 ± .035	.719 ± .023	.860 ± .075	.380 ± .55	.350 ± .015	.349 ± .012
MADNESS	Accuracy	**.938 ± .025**	.927 ± .026	.922 ± .032	.935 ± .012	.457 ± .018	.474 ± .067	.541 ± .067
	NMI	**.861 ± .030**	.853 ± .062	.857 ± .043	.864 ± .036	.346 ± .084	.404 ± .423	.230 ± .442

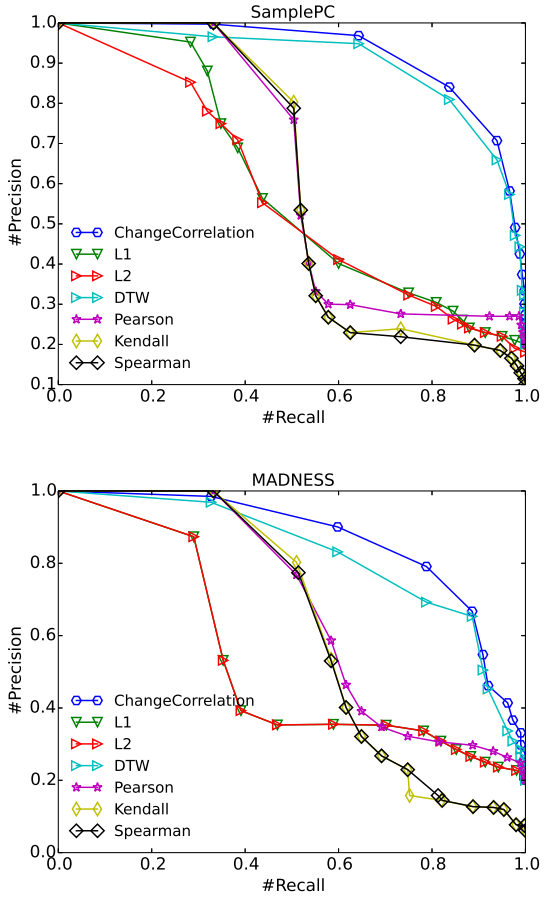

Figure 9: Precision Recall Curve (higher is better). We compare Change based correlation coefficient with other methods.

[20] W. Pirie. Spearman rank correlation coefficient. *Encyclopedia of statistical sciences*, 1988.

[21] D. M. Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness & correlation. *JMLT*, 2(1):37–63, 2011.

[22] L. Rabiner and B.-H. Juang. Fundamentals of speech recognition. 1993.

[23] B. Rosner. *Fundamentals of biostatistics*. Cengage Learning, 2010.

[24] M. Sonka, V. Hlavac, and R. Boyle. *Image processing, analysis, and machine vision*. Cengage Learning, 2014.

[25] C. Sun, H. Zhang, J.-G. Lou, H. Zhang, Q. Wang, D. Zhang, and S.-C. Khoo. Querying sequential software engineering data. In *Proceedings of the 22nd ACM SIGSOFT International Symposium on Foundations of Software Engineering*, pages 700–710. ACM, 2014.

[26] N. R. Tallent and J. M. Mellor-Crummey. Effective performance measurement and analysis of multithreaded applications. In *ACM Sigplan Notices*, volume 44, pages 229–240. ACM, 2009.

[27] L. P. Tilley et al. *Essentials of canine and feline electrocardiography*. CV Mosby., 1979.

[28] D. Wu, Y. Ke, J. X. Yu, S. Y. Philip, and L. Chen. Detecting leaders from correlated time series. In *Database Systems for Advanced Applications*, pages 352–367. Springer, 2010.

[29] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(1):34–58, 2002.

[30] Y. Zhu and D. Shasha. Statstream: Statistical monitoring of thousands of data streams in real time. In *VLDB*, pages 358–369. VLDB Endowment, 2002.