Large-Scale Correlation of Heterogeneous Time Series

Chen Luo, Lai Wei, Anshumali Shrivastava, John Mellor-Crummey
Department of Computer Science
Rice University
Houston, TX, USA
{cl67, Lai.Wei, anshumali, johnmc}@rice.edu

ABSTRACT

Measuring the correlation (or similarity) between different time series is a basic prerequisite for mining time series data. Traditional measures, such as Dynamic Time Warping (DTW), relies on the notion that two time series are similar if their have a lot of similar sub-patterns. However, there are many real-world scenarios where time series, having different overall patterns (e.g. Periodical, Linear, random, etc.), are still of common interest because they all change their behavior simultaneously. However, despite the importance of heterogenous time data, there has been little prior work addressing the correlation between two types of heterogeneous time series data.

In this paper, we propose a change based correlation measure, for mining heterogeneous time series. Our measure is ideally suited for heterogenous time series data commonly found in medical and high performance computing (HPC) domain. Furthermore, our measure admits locality sensitive hashing (LSH) scheme, leading to efficient sub-linear search and clustering algorithms. Existing of LHS makes our proposed measure ideal and practical for massive datasets. Rigorous experimental evaluations over **how many?** real and synthetic datasets clearly demonstrates the significant effectiveness and efficiency of our proposal over popular measures.

Categories and Subject Descriptors

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

General Terms

Application

Keywords

Correlation, Time Series, Hashing Learning

1. INTRODUCTION

SIGKDD '16 San Francisco, California USA

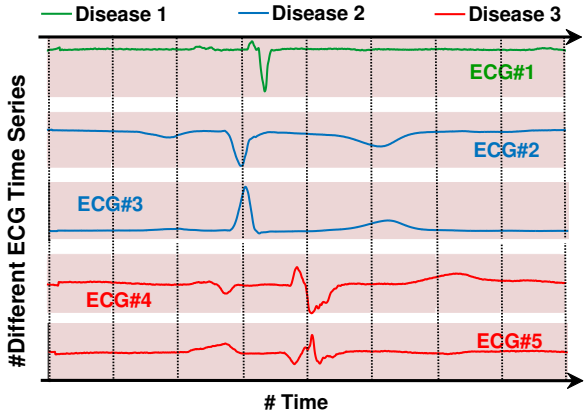

Figure 1: Five ECG time-series corresponding to three different diseases. Different color denotes different diseases.

The focus of this paper will be on the time series data. A time series is defined as a sequence of values $\{s_1, s_2, ..., s_m\}$ associated with timestamps $\{t(s_1), t(s_2), ..., t(s_m)\}$, which typically has the relationship of $t(s_i) = t(s_{i-1}) + \tau$, where τ is the sampling interval, and m is the number of points in the time series.

Time series mining is ubiquitous in data driven applications including robotics, medicine [25, 3], speech [29], object detection in vision [37, 32], High Performance Computing (HPC) and system failure diagnosis [19, 33], Earth Science [23], Finance [8], etc.

Due to its pervasive presence, time series mining have received significant attention in recent years. A common prerequisite for data mining algorithms, such as clustering, search, classification and regression, etc., is a measure of correlation (or similarity). Dynamic Time Warping (DTW) is widely accepted, and arguably the most popular, measure of similarity (or correlation) for time series data in general [30, 24, 5].

All existing measures over time series, including the popular DTW, relies on the notion that two time series S_1 and S_2 are similar, if there is a long enough, similarly behaving, subsequence common between them. This is a very reasonable notion which is also the desired nature of the similarity function in many applications. However, there are plenty of real-world problems where the similarity of interest has a very different notion. Despite their importance, there has been little previous work addressing such scenarios. To sig-

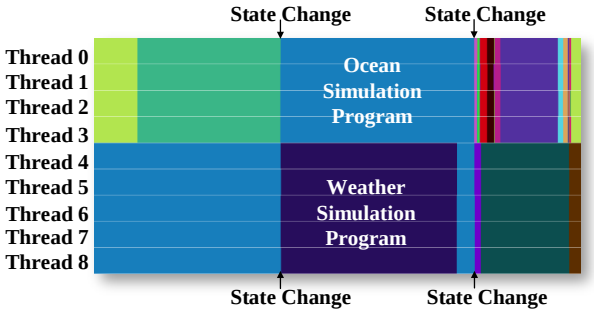

Figure 2: Snapshot of HPCToolket Traceviewer of an environment simulation task. Each line represent one thread, **Thread0** to **Thread3** belongs to a Ocean Simulation Program, and **Thread4** to **Thread8** belongs to a Weather Simulation Program. Different Color in each thread denotes the "Call Path" state.

nify their importance, we provide two motivating real-world examples:

ECG (Electrocardiogram) data

Electrocardiography [13] (ECG or EKG*) generates time series data, where each time series is the electrical activity of the heart over a period of time. This activity is measured using electrodes placed on a patient's body, which detect the tiny electrical changes on the skin that arise from the depolarization of the heart muscle during each heartbeat. The tiny electrical changes occurs in different rhythm corresponds to different heart physical states or different diseases. This depolarization if usually different for different patient. The characteristic of a disease is typically the time of sudden change in the electrical signal, i.e., if two ECG often have tiny electrical changes at the same time, they indicate same disease or same heart physical state [20].

As an illustration, Fig. 1 shows five ECG time series from a real data (Provided by UCR time series Archive [6]). These five ECG comes from three different diseases denoted by color green, blue and red. The DTW distance between ECG1 and ECG2 is 0.2, while the distance between ECG2 and ECG3 is 0.5. This is clearly not correct. Time series corresponding to same disease, for this ECG data, should have closer distance compared to time series corresponding to different diseases.

It is not surprising that we should not use the traditional measure of similarity for this task. This because that the similarity of ECG time series data correspond to the change rhythm, not correspond to the point to point similarity as evident from the Figure 1. Since the ECG time series data have different change patterns (e.g. Increase-Decrease, Decrease-Increase, or a Sudden wave, etc.). These different change patterns can effect the performance of point to point similarity measures.

On the other hand, By using the similarity in this work, the correlation between ECG1 and ECG2 is 1.00, and the similarity between ECG1 and ECG3 is 0.25, which clearly agrees with the gold standard labels.

High-Performance Computing (HPC) Logs: 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]. High-performance computer can generate over billions of threads during running. Automatically analysing and monitoring the HPC has recently grabbed significant attention by the HPC researchers [21, 34].

HPCToolkit[1] can be used to profile different threads from High performance computers. Information from each thread can be represented as a time series. The value of the time series denote the call path [1] state of the thread.

Fig. 2 shows a example of nine threads coming from an Environment simulation task. Here, Thread0 to Thread3 belongs to a Ocean Simulation Program, and Thread4 to Thread8 belongs to a Weather Simulation Program. Different Color in each thread denotes the "Call Path" state.

In this simulation task, the Ocean Simulation Program and the Weather Simulation Program are two correlated programs. These two programs need to exchange data during the environment simulation task executing. And once they communicate and exchange data with each other, their call path state will also change (As showed in Fig.2).

The DTW distance between Thread0 and Thread 4 is 0.8, which means they have large distance and does not correlate with each other. This is obvious contradict with the ground truth that they are correlated.

It is obvious that the correlation between these two programs can not be detected using DTW distance. Because threads in the different program always have different states and point to point similarity measures (DTW, etc.) will regard these state difference as large distance.

On the other hand, by using our method, we can find a 0.6 correlation coefficient with each other. This result clearly agrees with the ground truth that these two programs correlated.

As showed in above examples, the existing point to point time series similarity measures (e.g. L1-Distance, L2-Distance [11], and DTW-Distance [24], etc) or correlation measures (e.g. Pearson Correlation [26], Kendall rank correlation [17], and Spearman's rank correlation [27], etc.) can not deal with time series similarity problem when time series have heterogeneous patterns. The reason is: for heterogeneous time series, the correlation information is often associated with the change (During a period of time) of time series, rather than a point-to-point relationship. We will introduce the related research of point to point based similarity measure in detail in Section 5.

As a result, in order to deal with heterogeneity properties of time series, we proposed a change based correlation coefficient. 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. In addition, our measure admits locality sensitive hashing (LSH) scheme, leading to efficient sub-linear search and clustering algorithms.

The contribution of this paper is listed as follow:

1. Motivated by real applications, we investigate the correlation problem between heterogeneous time series. 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. By taking the advantage of this

[1]http://hpctoolkit.org/

correlation coefficient, we can use proposed measure under LSH scheme thus can do searching and mining on large scale time series datasets.

3. The experiments on Synthetic data sets and real-world data sets 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

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

2.1 Preliminary Definition

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.*

In most real world time series mining problems, the latency of a change in the time series is ubiquitous. For example, in a High-performance computer, a thread change of one stage may take some time to causal the other correlated state. And also for the ECG time series, the same disease may have slightly different latency in different bodies.

As a result, in this work, we allow the tiny difference between time stamps and define the soft equality of two timestamps as follow:

DEFINITION 4. *Given two time stamp t_1 and t_2, then $t_1 = t_2$ if and only if:*

$$t_1 - t_2 < \sigma \quad (1)$$

where, σ is a small time interval, and the value of σ depends on the real world problems.

2.2 Change based Time Series Correlation

As we mentioned before, in this work, we focus on using the change information to evaluate the correlation between two time-series. Before we formally define the change based correlation, we provide two intuitions of how two time-series can be correlated based on the change:

- If two time-series often change at the similar time, then they may correlate with each other. For example, the ECG time series showed in Fig. 1. ECG4 and ECG5 are corresponding to the same disease, so they are regarded correlated. On the other hand, ECG4 and ECG5 both have two electronic changes. And, from Fig. 1 we can see that the first change of each time series happens at the similar time, and the second change of each time series also happens at the similar time. So, they always have electronic changes at the similar time.

- How often does two time-series change at similar time denotes whether they are highly correlated or weakly correlated? In other words, if most of the time when one time-series changes, the other time series also changes, they may have a high correlation with each other. On the other hand, if just a few times that two time-series both changes, they may have a weak correlation. For example, in Fig.2, we can see that the threads in the same program change state all at the same time, this denotes that the threads within the same program are highly correlated. On the other hand, the thread from different programs changes states not always at the same time. This means threads from different programs may have weak correlation with each other.

Based on the two intuitions above, we define the correlation of this work as follow:

DEFINITION 5 (CHANGE BASED CORRELATION). *Suppose we have two time series: $X = (x_1, x_2, ..., x_m), Y = (y_1, y_2, ..., y_m)$, 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, the change based correlation coefficient is defined as the Jaccard distance [11] between C_X and C_Y.*

$$\rho(X,Y)^{ChangeCorrelation} = J(C_X, C_Y) = \frac{|C_X \cap C_Y|}{|C_X \cup C_Y|} \quad (2)$$

3. THE APPROACH

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

3.1 Change Based Correlation Evaluation Framework

The Change Based Correlation Coefficient can be calculated following the framework in Fig. 3. Given a time series, first, we 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 subseries, and 0 if not. We will introduce the change based information extraction in details in the following section.

After obtaining the change information of the time series, we then calculate the Jaccard similarity[11] coefficient between each other. As we defined before (Section 2), if two time series often change at the same time, they may have correlation with each other. So, here, how often denotes the value of the correlation. In other words, how many 1 do two time series both have. This is directly the Jaccard Distance. So, the Jaccard similarity between each bit-stream will be the correlation coefficient between these two time series.

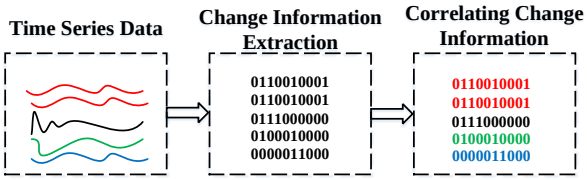

Figure 3: Overview of the Framework

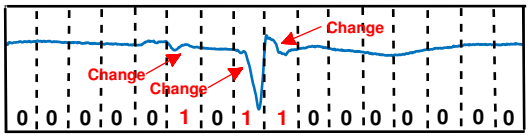

Figure 4: Change Information Extraction

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). The idea of extracting the change information is showed in Fig.4.

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

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:

$$b^l = \begin{cases} 1 & Have\ change\ in\ l \\ 0 & No\ change\ in\ l \end{cases} \tag{3}$$

As showed in Equ.3, 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 [18, 4] 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* [9]. We use the Two Sample t-test [22] method to solve this problem:

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

$$t_{score} = \frac{\overline{l^j_{Front}} - \overline{l^j_{Rear}}}{\sigma_p \sqrt{2/k}} \tag{4}$$

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:

$$\sigma_p = \frac{(k-1)\sigma^2_{l^j_{Front}} + (k-1)\sigma^2_{l^j_{Rear}}}{k-1} \tag{5}$$

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

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{6}$$

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 Search using LSH

In many the real world problems, the scale of time series data often huge [30]. Mining such huge number of time series is a big challenge for us.

In this work, by taking the advantage of the change based coefficient, we propose to use Locality Sensitive Hashing to do large scale search in time series data.

The change information of a time series is a bit-stream (Section 3.3). We regard it as the hashing code of the original time series.

As ar result, we can directly use LSH (Locality Sensitive Hashing) search algorithm to speed up the searching process. [14]

3.6 Connection Between proposed framework and Change based correlation definition

In this subsection, we briefly discuss about the connection between the proposed framework and the definition.

The proposed framework can detect change based correlation because of the following reasons: Recall the definition

of time stamps showed in Section 2, the equality of two time stamp is defined within a toleration of a small time interval σ. In other world, the change point in such small interval can be regarded as the same change. As a result, the if two time series both have change point at this small interval, that means these two time series have the same change point. So, the change information extraction can obtain all the change point information of the time series.

The Jaccard Similarity Coefficient used in our method just match the jaccard distance in the change based correlation coefficient.

Because of the above reasons, our method can detect the correlation just as defined.

3.7 Discuss about Sub-series Length σ

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

In some cases, the value of σ can be selected based on domain knowledge and experiments. In the experiment of this work, all the sub-series length are selected based on the domain knowledge.

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 our previous research of Correlating Event with time series [19], we can auto-select the sub-series length for a time series based on the autocorrelation function [10] of the time series. Given a time series $S = (s_1, s_2, ..., s_n)$, the autocorrelation is showed as follow:

$$R(l) = E(s_i * s_{i-l}). \tag{7}$$

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 [10]. Therefore, our length σ can be assigned as the value of the first peak to include the significant signal of the time series. For more detail of this selection method, please refer [19].

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, and several real world data sets.

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 [11], and DTW-Distance [24]. And for the three statistic correlation algorithm, we choose Pearson correlation coefficient[7], two ranking based correlation coefficients: The Kendall rank correlation [17], and Spearman's rank correlation [27].

In the rest of this subsection, we briefly introduce the three similarity measures and the three Correlation measures.

4.1.1 Similarity Measures

Given a two time series $X = (x_1, x_2, ..., x_m), Y = (y_1, y_2, ..., y_m)$, where m is the length of the time series.

The L1-distance is denoted as follow:

$$L_1(X, Y) = \sum_1^m |x_i - y_i| \tag{8}$$

The L2-distance is denoted as follow:

$$L_2(X, Y) = \sqrt{\sum_1^m |x_i - y_i|^2} \tag{9}$$

The third similarity measure we used is DTW distance [24], which is a famous time series similarity measure.

In order to introduce the DTW distance, we firstly 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 cost:

$$DTW(X, Y) = \min \sqrt{\sum_{k=1}^K w_k} \tag{10}$$

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 [26], Kendall rank correlation [17], and Spearman's rank correlation [27].

The Pearson correlation coefficient, can be calculated as follow:

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

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 [17] is defined as follow:

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

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 [27] is defined as follow:

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

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

4.2 Evaluation Strategy

In order to evaluate both the correctness and efficiency of our method, we design two tasks using our measure and other measures: Time Series Clustering Task, and Time Series Top-k Search Task. In the rest of this subsection, we will brief introduce these two tasks in details.

4.2.1 Clustering Task

In order to evaluate the performance of our correlation coefficient, we design a clustering task. In this task, we use

the Hierarchical Clustering [11] to evaluate the performance. Hierarchical Clustering is very sensitive to the distance measure. So, the clustering performance can directly illustrate the effectiveness of our method compared with other methods.

Two evaluation methods are used for testing the clustering result: Accuracy [11], which is calculated as the percentage of target objects clustered into the correct clusters. The clustering accuracy (r) is defined as

$$r = \frac{\sum_{i=1}^{k} a_i}{n}$$

where a_i is the number of data objects occurring both in ith cluster and its corresponding true class, and n is the number of time series.

Normalized Mutual Information (NMI) [11], which is one of the most popular evaluation methods to evaluate the quality of clustering results. The normalized mutual information of two clustering result is defined as follow:

$$NMI(X,Y) = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$$

where X and Y are vectors containing cluster labels for all the target objects, $I(X;Y)$ denotes the mutual information[11] of these two vector, $H(X)$ and $H(Y)$ are the marginal entropies[11]. Both accuracy and NMI are in the range of 0 to 1, and a higher value indicates a better clustering result in terms of ground truth.

4.2.2 Top-K Searching Task

In this work, we propose to use LSH scheme to do fast search. So, we design a top-k search task for by using different measures.

For the change based correlation measure, we use our proposed LSH search scheme. For the other similarity or correlation measures. We use naive based search. The step of naive based method is: (1) calculate all the distance between the query item and the items in the data set. (2) Ranking all the items based on the distance calculated in first step. (2) Return top k items. Naive based search can get the best search result for each measure.

We compute precision and recall [28] to evaluate all the top-k search result. 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:

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

and the recall can be calculated as follow:

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

4.3 Effectiveness Study on Synthetic Dataset

In this section, we introduce the experiment on the synthetic Dataset. We first brief describe the synthetic dataset, and then show the experimental result of each tasks on this data sets by using different measures.

4.3.1 The Synthetic Dataset

Table 1: **Change Types in the 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

Table 2: 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

Synthetic data set is very useful for evaluating algorithms and functions for data mining and machine learning models[11]. The steps of generating this synthetic data is showed as follow:

Step 1 randomly generate 10000 time series with two patterns of time series: (1) Periodical Pattern, (2) Linear Pattern. The length of each time series is 5000.

Step 2 Add white noise and correlated noise randomly [11] on each time series.

Step 3 Random shuffle all the 10000 time series and Divide these 10000 time series into 5 groups.

Step 4 Add change randomly into 5 groups, within same group, the change times are same, with different groups the change time are different. Seven different types of changes are added randomly into each time series, the seven change types are showed in Table.1.

After obtaining these 10000 time series with five groups. In order to test both the efficiency and effectiveness of change based correlation coefficient. We choose four sub-set from these time series. We make the data set size from small to large, and also the time series length from short to long. The four sub dataset is showed in Table.2.

4.3.2 Clustering task on Synthetic Data

The clustering task result on these four synthetic data are showed in Table.3.

From Table.3, 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.

Table 3: 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

Table 4: Query Time for Top-10 Search

Measure	Data Set			
	T0[2]	T1	T2	T3
Proposed	0.004	0.01	0.0033	0.08
L1	0.01	0.1	0.13	0.85
L2	0.02	0.15	0.15	1.05
DTW	0.7	107.59	15.82	0.20
Pearson	0.22	0.29	2.04	0.85
Kendall	2.80	55.39	26.53	2.07
Spearman	0.4	1.81	3.87	1.18

4.3.3 Top-k Searching task on Synthetic Data

The plots for all the four datasets are shown in Figure.5. 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.

Table.4 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.

4.4 Effectiveness Study on Real Datasets

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

4.4.1 Electrocardiogram Data set

The first real world dataset is ECG (Electrocardiogram) time series data set. This data set comes from the the UCR time series Data set Archive [6]. We choose four ECG data set there as showed in Table. 5. And the ground truth comes from the UCR data set itself.

For the clustering task, we use the Hierarchical Clustering [11] to evaluate the performance as before. From Table.6, we can see that, for the change based correlation can obtain more accuracy results in these four ECG data set compared with both the similarity method and the correlation coefficient. So, the result of clustering show the effectiveness of our coefficient.

For the Top-K searching task. The plots for all the four ECG dataset 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. This demonstrate the effectiveness of change-based correlation coefficient and the corresponding LSH search algorithm.

Table.7 shows the Execution time of the top-k searching task on each ECG data set. From the result, we can see that change based correlation performed much faster than other algorithms.

4.4.2 HPC Thread Time Series Data set

The second real world dataset is collected from the HPC-ToolKit during three HPC task running.

Single PC

Single PC is a small toy execution task that execute on a single PC and small number of threads (20). In this datasets, there are two types of threads: Main Thread and Child Thread. We use the label of Main Thread and Child Thread as ground truth information.

MADNESS

MADNESS[3] is a high level software sofware environment for Scientific Simulation. We use this environment to simulation several Scientific calculations, and generate 264 threads. The same of Single PC dataset, there are two types of thread: Main Thread and Child Thread. We also use them as ground truth information.

Earth Environment Simulation

This data set is collected from an earth Environment Simulation task. In this task, there are a several of simulation programs, including Oceans Simulation Programs, Weather simulation Programs, and several other simulation programs. The Oceans Simulation Program, and the Weather simulation Program are correlated with each other due to high frequency data transformation and communication. So the ground truth in this data is defined that programs with high frequency communications are having the same label. ...

We perform both clustering task and top-k search task on these three data set.

From Table.10, we can see that, for the change based correlation can obtain more accuracy results in the two HPC

[3]https://en.wikipedia.org/wiki/MADNESS

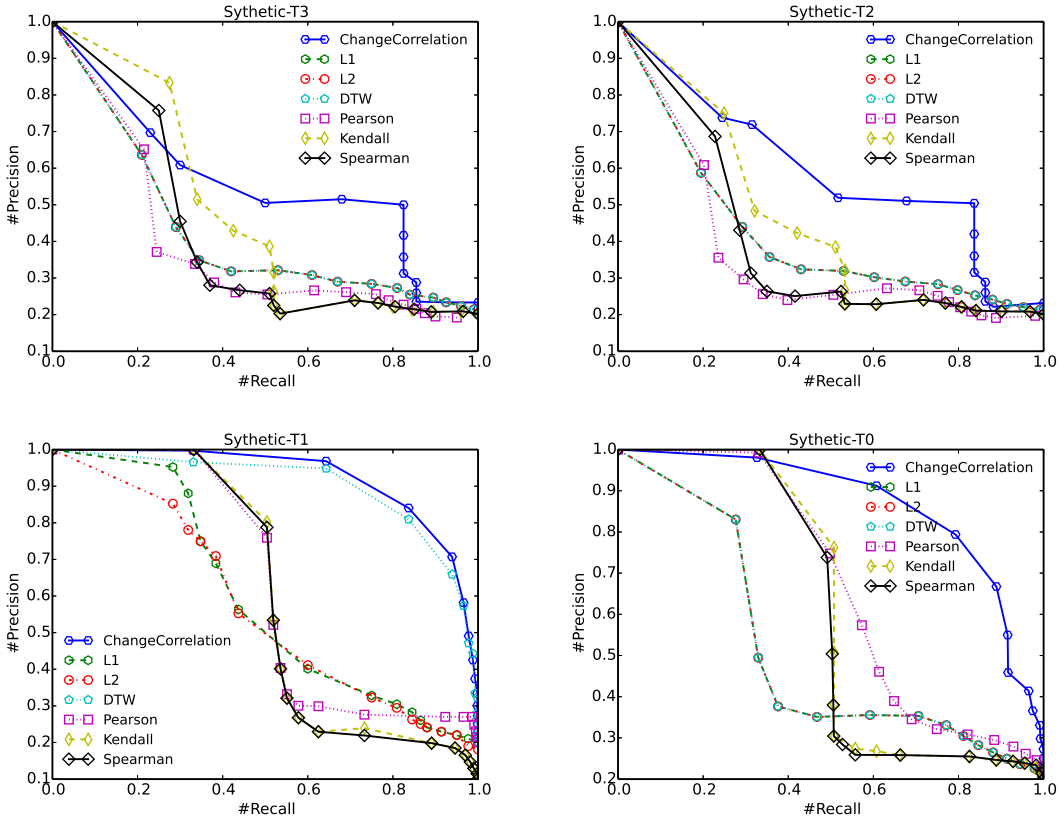

Figure 5: Precision Recall Curve for Different Algorithms

Table 5: Summary of the Four ECG Data Set

Data Set	Data Size	Time Series Length	Class Numver
CinC_ECG_torso	1380	1639	4
ECGFiveDays	861	136	2
TwoLeadECG	1139	82	2
ECG5000	4500	140	5

Thread Time series data set compared with both the similarity method and the correlation coefficient. While we can see that DTW method can also get high accuracy, this because in data set, most of the thread in same class often in same state. However, as we said before, in some cases, thread in same class, the state can be different.

The precision recall curve for the HPC Thread TIme Series data is showed in Figure 7. We can clearly see that our proposed Change-based Correlation method gives higher precision recall curves than other similarity and correlation methods. Also, DTW method can get good result too. In addition the results are consistent across datasets. This demonstrate the effectiveness of change-based correlation coefficient and the corresponding LSH scheme.

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 [15]. Pearson Correlation [7] is a basic correlation measure between time series, which has been widely used in practice [38]. 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 [35], 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 [31] 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

Table 6: 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 7: Query Time for Top-10 Search

Measure	Data Set			
	CinCECGtorso	ECG5000	TwoLeadECG	ECGFiveDays
Proposed	0.005	0.02	0.003	0.01
L1	0.04	0.07	0.01	0.028
L2	0.05	0.078	0.11	0.030
DTW	12.55	0.59	0.025	0.20
Pearson	0.31	0.15	0.21	0.85
Kendall	10.67	0.39	0.39	2.07
Spearman	0.82	0.23	0.28	1.18

Table 8: Summary of the HPC Time Series Data Set

Data Set	Data Size	Time Series Length
Single PC	24	4096
MADNESS	264	32768
Environment Simulation	312	16384

Table 9: Query Time for Top-10 Search

Measure	Data Set		
	SinglePC	MADNESS	Earth
Proposed	0.02	0.16	0.10
L1	0.18	1.02	0.53
L2	0.48	5.27	2.79
DTW	21.83	2537.21	898.31
Pearson	0.35	2.45	1.20
Kendall	9.25	1324.26	536.32
Spearman	1.25	315.45	112.56

events, the above rank based algorithms cannot be directly used in our scenario.

5.2 Change Point Detection

The problem of change detection has been studied for a long time, and various methods such as CUSUM (cumulated summation) [2], wavelet analysis [16], inflection point search [12], and Gaussian mixtures [36] have been proposed. These algorithms can be used in our method. However, our provided method can quickly detect the boolean problem of whether there is a change in the time period. We do not need to find the time series change point very accurate.

6. CONCLUSION AND FUTURE WORKS

Calculating the correlation between different time series is an important data mining task. By using correlation method, we can find hidden relationships between different time series. However, in most real world problems, time series often have very different patterns (e.g. Periodical, Linear, random, etc.). And there also correlations between such heterogeneous time series, and the correlation between these time series is also a very import problem. 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 top-k searching problems. 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*

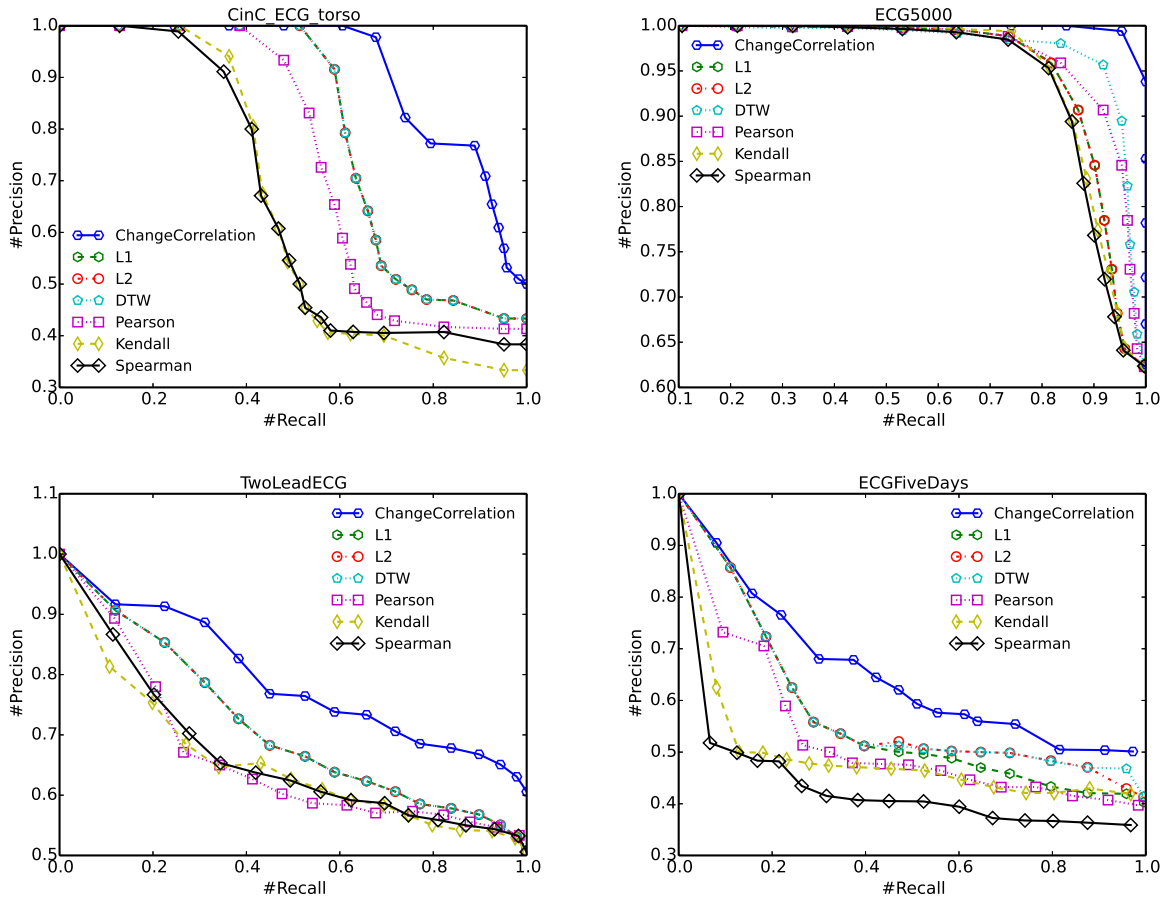

Figure 6: Precision Recall Curve (higher is better). We compare Change based correlation coefficient with other methods.

Computation: Practice and Experience, 22(6):685–701, 2010.

[2] M. Basseville, I. V. Nikiforov, et al. *Detection of abrupt changes: theory and application*, volume 104. Prentice Hall Englewood Cliffs, 1993.

[3] J. P. Caraça-Valente and I. López-Chavarrías. Discovering similar patterns in time series. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 497–505. ACM, 2000.

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

[5] Y. Chen, B. Hu, E. Keogh, and G. E. Batista. Dtw-d: time series semi-supervised learning from a single example. In *KDD*, pages 383–391. ACM, 2013.

[6] 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/`.

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

[8] C. W. J. Granger and P. Newbold. *Forecasting economic time series*. Academic Press, 2014.

[9] 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.

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

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

[12] S. Hirano and S. Tsumoto. Mining similar temporal patterns in long time-series data and its application to medicine. In *Data Mining, 2002. ICDM 2003. Proceedings. 2002 IEEE International Conference on*, pages 219–226. IEEE, 2002.

[13] 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.

[14] 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.

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

[16] S. Kadambe and G. F. Boudreaux-Bartels. Application of the wavelet transform for pitch

Table 10: 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
Simulation	Accuracy	**.921 ± .021**	.802 ± .026	.420 ± .032	.435 ± .012	.857 ± .018	.464 ± .067	.531 ± .067
	NMI	**.831 ± .030**	.803 ± .062	.537 ± .043	.504 ± .036	.786 ± .084	.202 ± .423	.330 ± .442

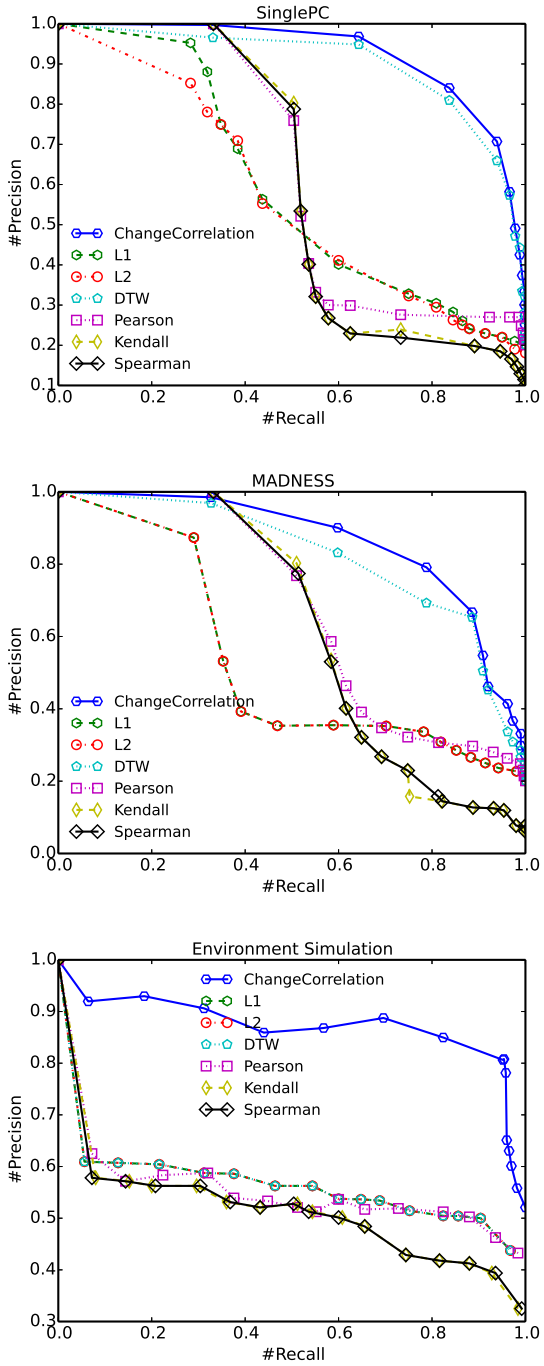

Figure 7: Precision Recall Curve (higher is better). We compare Change based correlation coefficient with other methods.

detection of speech signals. *IEEE Transactions on Information Theory*, 38(2):917–924, 1992.

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

[18] 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.

[19] 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.

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

[21] 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.

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

[23] M. Mudelsee. *Climate time series analysis*. Springer, 2013.

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

[25] T. Oates, M. D. Schmill, and P. R. Cohen. A method for clustering the experiences of a mobile robot that accords with human judgments. In *AAAI/IAAI*, pages 846–851, 2000.

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

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

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

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

[30] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270. ACM, 2012.

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

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

[33] 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.

[34] 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.

[35] 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.

[36] K. Yamanishi and J.-i. Takeuchi. A unifying framework for detecting outliers and change points from non-stationary time series data. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 676–681. ACM, 2002.

[37] 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.

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