http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 12 (19): 4906-4910, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.4906.4910

A Novel Never-ending Uncertain Top-k Discord Detection Method

[1]Yan Qiuyan and [1]Chen Xiongtao
[1]School of Computer Science and Technology,
China University of Mining and Technology, China

Abstract: Time series discords are subsequences that are maximally different to all other time series subsequences of a longer time series. Discord Detection is widely used in time series applications. We observed that the discord position are often changed when noise data interfere with the time series. This phenomenon is produced because the traditional method cannot concern the factor that noise data infect the original data's distribution. In this study, we opposed a novel method which combined top-k discord detection with uncertain ranking to achieve uncertain top-k discord detection. Through transforming the discord score interval to satisfied with Gaussian distribution, the new method can ranking series data with arbitrary distribution. Finally, we demonstrate a comprehensive experimental study to verify the effectiveness and efficiency of the proposed approach.

Key words: Discord detection, continuous uncertain data, top-k discord, time series

INTRODUCTION

Time series discords are subsequences that are maximally different to all other time series subsequences of a longer time series. Thus, time series discords capture the sense of the most unusual subsequence within a series. Discords detection have many applications in medicine, finance, biology, engineering and industry. The Brute Force Discord Detection (BFDD) algorithm suggested by Keogh *et al.* (2005) is an exhaustive search algorithm that requires time complexity O (n^2) to detect discords. Typical discord detection methods, such as BFDD, HOT SAX (Keogh *et al.*, 2005), iSAX (Camerra *et al.*, 2010) all compute the maximally different subsequences for determined data. However, in certain circumstances, the query top-k discords of the time series are more valuable. For example, coalmine gas concentration monitoring is a typical real-time monitoring application for top-k discord detection. When the kth discord value of gas concentration data is greater than the threshold, the monitoring system must immediately produce an alarm to avoid gas explosion accidents. On the other hand, when noise data interfere with the time series, a greater noise variance corresponds to a greater impact on real data and the detected discord position is changed accordingly.

RELATED WORKS

Traditional discord direction methods (i.e., BFDD, HOT SAX, iSAX) can manage only accurate or determined data. When such methods are directly applied to time series with noise data, the discord position obtains bias.

SAX-based discord detection method (Khanh and Anh, 2012; Buu and Anh, 2011) applied Piecewise Aggregate Approximation (PAA) technology, which running once on the whole data. So the array structure of SAX is designed according to the total data number. In the online monitoring system, the data is coming in a never-ending fission, so the SAX method can't be effective.

For top-k ranking on the sequence of discord position, the top-k ranking method for uncertain data (Soliman and Ilyas, 2009) is either suitable for uniformly distributed data or discrete data (Li *et al.*, 2011). However, in reality, most time series data are not uniformly distributed and the data distribution cannot be determined in advance. Thus, existing methods cannot solve the arbitrary distribution problem of top-k discord detection. The rest of this study is organized as follows: In Section 3, we presents the Never-ending discord detection method NE_TKDD and the improved Continuous Uncertain top-k algorithm for arbitrary

Corresponding Author: Yan Qiuyan, School of Computer Science and Technology, China University of Mining and Technology, China

distribution CU_TK and the last the combined algorithm CU_TKDD are specified. Section 4 presents the results of the correlation algorithm experiment.

UNCERTAIN TOP-K DISCORD DETECTION FOR ARBITRARY DISTRIBUTION SERIES DATA

On the basic conception of above analysis, our study proposes an never-ending uncertain top-k discord detection method for continuous time series with unknown distribution.

Firstly, A SAX-based deterministic top-k discord detection method is improved facing the never-ending environment to satisfy the online monitoring systems which applies top-k discord detection technology most frequently. Secondly, the score interval of discord is transformed and the approximate uniform distribution function can be obtained. So we can apply MCMC (Soliman and Ilyas, 2009) method when score density function is unknown. Thirdly, the never-ending determined top-k discord detection method combined with uncertain top-k algorithm for arbitrary distribution, getting the final uncertain top-k discord detection method. In the next sections, we formulate the above three contents separately.

NE_TKDD method: Present study is a preliminary attempt to detection top-k discord for arbitrary distribution series data and our idea of resolving this problem is that we can first detect top-m discord sets DS: {p1, …, pm}, (m≥k) using deterministic algorithm and then ranking the discord sets according to uncertain top-k method. Based on this motivation, we only select the HOT SAX method to verify the feasibility of our method because HOT SAX is widely used in many application fields. The other discord detection method with higher efficiency will be discussed in our future work. The main optimized content of NE_TKDD (Never-Ending Top-K Discord Detection) Method consists of two aspects: A combined two dimensional (2d) table and a priority queue structure, this structure replaces the fixed-length sequence-based 2d table in SAX and Dict-Tree, thereby solving the storage problem of never-ending data.

HOT SAX method stores the number and starting position of the string in a 2d table. The string is stored in the table according to the starting position of the character sequence. When the volume of data set is already known, the dimension of the 2d table is constant. However, when data flow is continuous and never-ending, the volume of 2d table cannot be increased accordingly. In this study, the original 2d table used in HOT SAX is replaced by two 2d table structures. The length of both 2d tables are w^a where a is the number of character sets, w is the length of the SAX string, a and w are all usually small constants, such as a = 3, w = 3. Thus, the updating complexity must be significantly decreased even if the data volume are infinite.

Discord score interval transformation: After improved the HOTSAX structure, we can use the determined discord detection method to get the top-k discord set DS: {p1, …, pm}, (m≥k). But this set is a result that didn't concern the data distribution with noise disturbance. Next step, we need to consider the uncertain top-k query from data set DS.

Similar to the deterministic data ranking algorithm, the score interval of discord record is the main element that influences the uncertain data ranking. In the time series of discord detection, the non-self match distance among the subsequences is defined as the discord score. When influenced by noise data, the distance between the subsequence calculations is affected and complied with a certain probability distribution. However, the distribution function is difficult to obtain, thus preventing the direct use of the uncertain top-k method for continuous data.

To solve this problem, s points are taken around the abnormal subsequences that have p_i starting positions (s < m, such as s = 5; m is the buffer size). The abnormal sub-interval $[p_i-s, p_i+s]$ is then obtained because the interval of 2s is small, thus suggesting that the abnormal values in this interval can be considered uniformly distributed. The density function of the sub-interval can be expressed as:

$$f(x)=\frac{1}{2s}$$

The 2s max and min distances of the subsequences are then considered the interval caps and collars. This interval is marked as a tuple. The referenced MCMC method, which is an uncertain data-sorting algorithm based on uniform distribution is then employed to perform the top-k sorting of the discord tuple. Following is the improved Continuous Uncertain Top-K algorithm for arbitrary distribution.

Algorithm: CU_TK
Input: top-k discord set DS: {p1, …, pm}, (m≥k)
Output: uncertain top-k tuple ranking list

begin
Randomly designate a branch of the tree as state S_0 and a position r, $1 \le r \le k$
While (t_r incompletely dominates $t_{(r+1)}$ and the location of [1, k] has not been selected) do begin the tuple in location r is marked as t_r. Swap t_r with $t_{(r+1)}$ to obtain the new state S_1.
Compare the size of the k prefix probability of S_0 and S_1, as well as Pr (S_0) and Pr (S_1)
If Pr (S_0) >Pr (S_1), accept state S_1 and mark S_1 as S_0, select another position r of [1, k];
Else continue to move t_r downward, r = r+1;
End while;
return the state S_0.

Combined uncertain top-k discord detection method: Until now, we e improved the determined HOTSAX-based top-k discord detection method to get top-k discord record set DS and transformed the discord score interval to query top-k discord record adapting to uncertain top-k MCMC method.

Algorithm: CU_TKDD:
Input: Time series dataset T
Output: Top-k Discord Set
Step 1: Call the O_TKDD algorithm and determine the top-k discord set. TD: $\{p_1, \ldots, p_m\}$, $(m \ge k)$, where p_i is the beginning of the sequence abnormality.
Step 2: Calculate the discord sequence score interval for each record in TD. The discord score intervals are then transformed.
Step 3: Compare the discord record scores to obtain the Hasse diagram, which expresses the dominant relationship.
Step 4: Call the CU_TK algorithm to obtain the top-k discord record set.

PERFORMANCE STUDY

Datasets and setup: We perform an extensive evaluation by using the following datasets:

- **Ma (Ma and Perkins, 2003):** This time series is formulated by the sine plot and has 1200 data points. We insert ten groups of noise to the series to detect anomaly subsequence
- **Input:** This time series is the gas concentration monitoring series data of electromagnetic radiation. The sampling interval is 0.5 sec. The total number of data is 7500, which contains the actual 10 gas abnormal positions

Detection accuracy: All experiments are performed on a machine with an Intel Core 2.53 GHz CPU and 2 GB RAM. We compared CU_TKDD with HOTSAX, MCMC+HOTSAX and NE_TKDD to evaluate the accuracy of CU_TKDD. Thus, we first call the HOTSAX method to compute the top-k discord before inserting noise to the datasets. We tag the result as discord set0, which express the deterministic results. Thereafter, we insert different noise to the four datasets and call the HOTSAX, NE_TKDD, MCMC+HOTSAX and CU_TKDD methods to compute the top-k discord. We tag the result as discord sets1 to sets4. We compute the accuracy rate by comparing the accuracy of NE_TKDD MCMC+HOTSAX and CU_TKDD after inserting noise with the deterministic results got by HOTSAX. For example, according to NE _TKDD, the accuracy rate is equal to (sets0-(sets2 n sets0))/sets0.

We test the effect of different white noise variances on HOTSAX, NE_TKDD, MCMC+HOTSAX and CU_TKDD. We find that the white noise variance significantly affect only the datasets that have flat tendencies, such as Ma data. However, for datasets with amplitudes that change dramatically, such as Input data, we must insert a large deviation noise to perceive the influence of noise data on the top-k ranking result.

Thus, we set the variance value from 0.05 to 2 for Ma and from 0.05 to 12 for Res and Input.

As shown in Fig. 1, for all datasets, HOTSAX accuracy decreases with increasing uncertainty, thus proving the necessity of uncertain data ranking. Generally speaking, MCMC+HOTSAX present a highest accuracy than all the other methods.The reason is what we have said that HOTSAX method can't adapt to the never-ending environment, when data volume is infinite, HOTSAX can't applied, thus the road map of MCMC+HOTSAX can't work also.

In order to process infinite series data, NE_TKDD method transformed the String storage structure of HOTSAX and appended the circular linked list. Thus NE_TKDD has a less accuracy than MCMC+HOTSAX but higher accuracy than HOTSAX when noise interfere is huge.

When the noise is small, MCMC+HOTSAX and HOTSAX has the similar accuracy, with the noise affect increasing, MCMC+HOTSAX and CU_TKDD method have the higher accuracy. On the whole, CU_TKDD method has the accuracy similar with MCMC+HOTSAX and higher than traditional HOTSAX method when taking into the noise interfering.

Efficiency evaluation of the CU_TKDD method: In this experiment, we test the running time for different datasets with different data numbers and k values. Figure 2 shows the different data number execution times when the k value increases (k = 3, 5, 7, 9). When the data umber is a fixed value, the running time of CU_TKDD increases linearly with the k value. When the data number varies, the running time varies because of the varying buffer size. We select a larger buffer size for a larger data size.

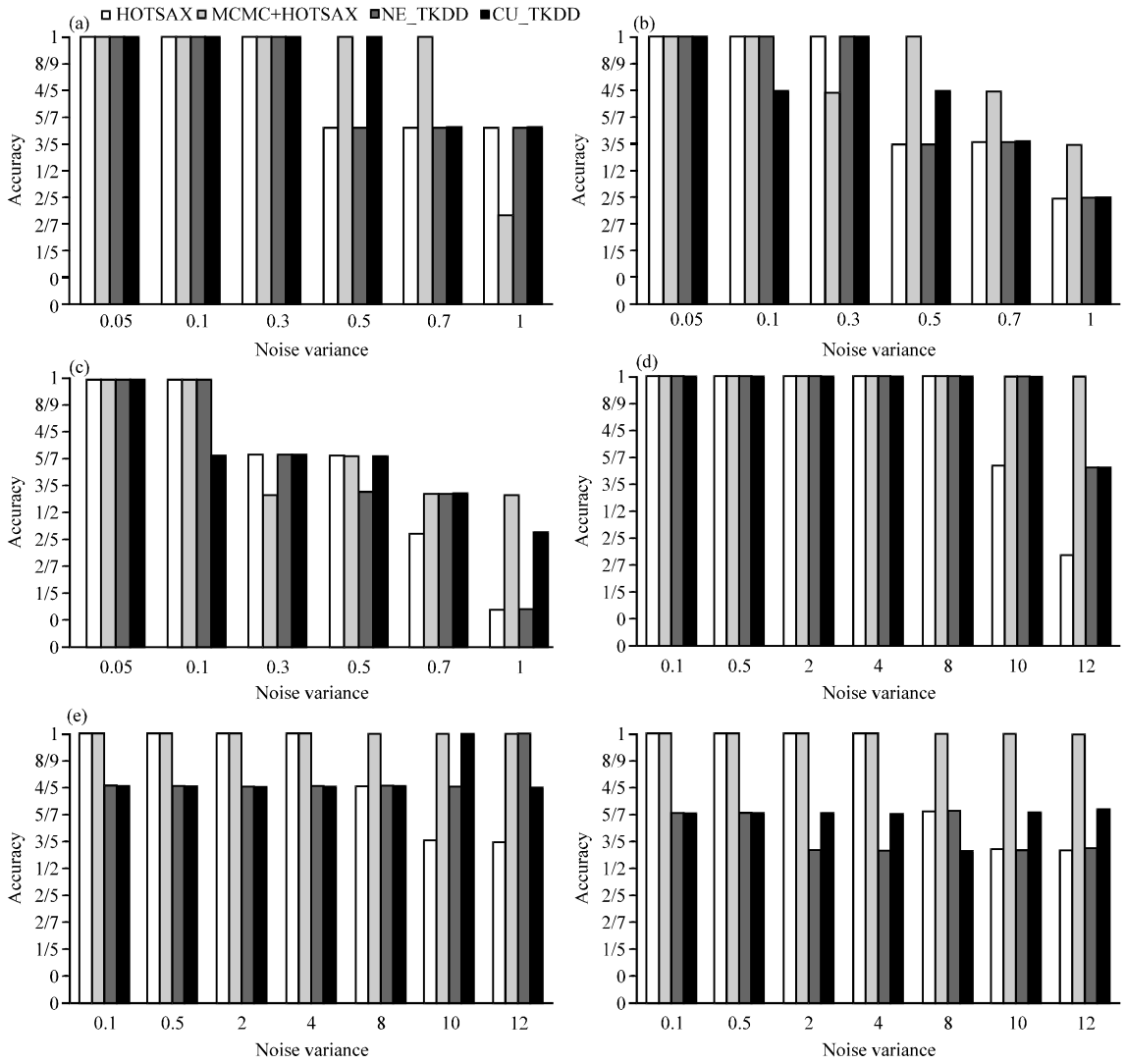

Fig. 1(a-f): Detection accuracy for all dataset, (a) k = 3 for Ma, (b) k = 5 for Ma, (c) k = 7 for Ma, (d) k = 3 for Input, (e) k = 5 for Input and (f) k = 7 for Input

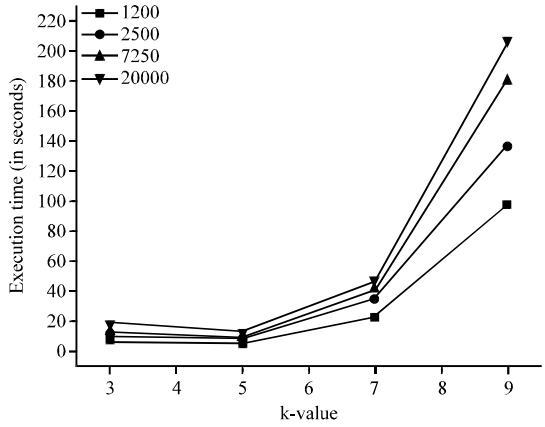

Fig. 2: Running time of the CU_TKDD with different k values

CONCLUSION

We propose a continuous uncertain time series top-k anomaly detection method with arbitrary distribution in the never-ending circumstance. Anomaly detection method with arbitrary distribution in the never-ending circumstance. The proposed method can detect discord records accurately with a faster response time than traditional method in a never-ending manner despite disturbance from noise data. The proposed method is evaluated by using real world and synthetic datasets to verify its efficiency and accuracy.

REFERENCES

Buu, H.T.Q. and D.T. Anh, 2011. Time series discord discovery based on iSAX symbolic representation. Proceedings of the 3rd International Conference on Knowledge and Systems Engineering, October 14-17, 2011, Hanoi, Vietnam, pp: 11-18.

Camerra, A., T. Palpanas, J. Shieh and E.J. Keogh, 2010. iSAX 2.0: Indexing and mining one billion time series. Proceedings of the IEEE International Conference on Data Mining, December 14-17, 2010, Sydney, Australia, pp: 58-67.

Keogh, E., J. Lin and A. Fu, 2005. HOT SAX: Efficiently finding the most unusual time series subsequence. Proceedings of the 5th IEEE International Conference on Data Mining, November 27-30, 2005, Houston, Texas, USA., pp: 226-233.

Khanh, N.D.K. and D.T. Anh, 2012. Time series discord discovery using WAT algorithm and iSAX representation. Proceedings of the 3rd Symposium on Information and Communication Technology, August 23-24, 2012, Ha Long, Vietnam, pp: 207-213.

Li, J., B. Saha and A. Deshpande, 2011. A unified approach to ranking in probabilistic databases. VLDB J., 20: 249-275.

Ma, J. and S. Perkins, 2003. Online novelty detection on temporal sequences. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003, Washington, DC., USA., pp: 613-618.

Soliman, M.A. and I.F. Ilyas, 2009. Ranking with uncertain scores. Proceedings of the IEEE International Conference on Data Engineering, March 29-April 2, 2009, Shanghai, China, pp: 317-328.