A WED Method for Evaluating the Performance of Change-Point Detection Algorithms

Jin-Peng Qi
College of Information Science & Technology,
Donghua University,
Shanghai, China
qipengkai@dhu.edu.cn

Ying Zhu
Hunter New England Health,
Royal North Shore Hospital,
New South Wales, Australia
ying.zhu@hnehealth.nsw.gov.au

Ping Zhang
Menzies Health Institute,
Griffith University,
Queensland, Australia
p.zhang@griffith.edu.au

Abstract— Change point detection (CPD) is to find the abrupt changes in a time series. Various computational algorithms have been developed for CPD. To compare the different CPD models, many performance metrics have been introduced to evaluate the algorithms. Each of the previous evaluation methods measures the different aspect of the methods. In this paper, a new weighted error distance (WED) method is proposed to evaluate the overall performance of a CPD model across multiple time series of different lengths. A concept of normalized error distance was introduced to allow comparison of the distances between an estimated change point position and the target change point among models that work on multiple time series. In this study, the WED metrics was applied on synthetic datasets with different sample sizes and variances to evaluate the different CPD models, including: Kolmogorov-Smirnov (KS), SSA and T algorithms. The test results showed the value of this WED method that contributes to the methodology for evaluating the performance of CPD models.

Keywords—change point detection, weighted error distance, WED, MWED

I. INTRODUCTION

Change point detection (CPD), or abrupt change detection, is the application of techniques to detect changes in properties of a time series. Detection of abrupt changes has been widely studied in many real-world problems, such as: atmospheric and financial analyses [1], fault detection in engineering systems [2, 3], climate change detection [4], and genetic time-series analyses [5]. The usage of this method to detect pattern changes in ECG and EEG signals may also be beneficial. This application would allow appropriate staff to be alerted of changes in a patient's medical situation and to provide on-time treatment [6, 7]. CPD models utilize the algorithms that cover the fields of data mining, statistics, and computer science, including parametric and nonparametric methods [8, 9, 10, 11]. Each CPD algorithm can be assessed from the aspect of detection accuracy, computational cost or whether it can be a real time detection.

Many performance metrics have been introduced to evaluate CPD algorithms based on the type of decisions they make [12]. Aminikhanghahi and Cook [13] reviewed the performance evaluation methods commonly used for CPD models. The evaluation can be based on a yes/no decision – whether the change point (CP) was detected within certain distance from the real change point. In this case, the CPD model can be treated as a binary classification model and can be evaluated with the usual measures, such as accuracy, sensitivity, specificity or ROC curve [14, 15]. For real applications, for example clinical decision makings, cut offs applied to the model outcomes can be adjusted to achieve different sensitivity and specificity [16]. However, when the difference in time between the detected/estimated change point (e-CP) and the actual CP represents the measure of CPD performance, then the evaluation of these algorithms is not as straightforward as for the binary classification. There is no single label against which the performance of the algorithm can be measured. A number of useful metrics take into account the distance between e-CP and actual CP to measure CPD method performance. These metrics include: mean absolute error (MAE), mean squared error (MSE), mean signed difference (MSD), root mean squared error (RMSE) and normalized root mean squared error (NRMSE). Of these, except NRMSE normalizes the unit size of the predicted value and facilitates more direct comparison of error between different datasets, the other methods measure only the absolute distances between e-CP and actual. However, even NRMSE does not count the difference between the situations when the e-CP is before the actual CP and when it is after the actual CP. It also fails to consider the relative position of the actual CP within the total length of the time series.

In this study, we introduced a concept of weighted error distance (WED) which can be interpreted as a normalized distance between the e-CP and actual, or target, change point (t-CP). A WED metrics is proposed to compare the overall performance of CPD models working across multiple time series of different lengths and the occurrence of t-CPs at different positions in the time series. The ability of this WED metric to evaluate different CPD models was tested on Kolmogorov-Smirnov (KS) [8, 11], SSA [17, 9] and T [18] algorithms that worked on the synthetic datasets in different sample sizes and different variances. The implementations of KS, SSA and T for CPD on time series were described in our previous studies [8, 19].

II. METHOD

Suppose a time series, $X = \{x_a \ldots x_c \ldots x_e\}$, starts from time point *a* and ends with time point *e* with a single change point. The time series *X* is composed of two adjacent segments, in which the former (left) part $\{x_a, \ldots, x_{c-1}\}$ and the latter (right) part $\{x_{c+1}, \ldots, x_e\}$ are divided by a predefined target change point, x_c. From a statistical point of view, we refer to the former (left) part as a positive area and the latter (right) part as a negative area. When applying a CPD to detect the change point within the time series, the e-CP can occur in either the positive or negative area. A few concepts are

This project is supported by National Natural Science Foundation of China (no. 61104154) and Specialized Research Fund for Natural Science Foundation of Shanghai (no. 16ZR1401300 and no.16ZR1401200).

introduced here before the WED metrics is presented to measure CPD model performance: true positive distance (TPD), positive error distance (PED), true negative distance (TND) and negative error distance (NED). As shown in Fig. 1, if the e-CP is located on the left side of the t-CP (positive area) the PED and TPD can be calculated, that is the distance from the e-CP to the t-CP and to the start point, respectively. Meanwhile, NED and TND are not applicable. Conversely, when the e-CP is on the right side of the t-CP (negative area), then TPD and PED do not exist. NED is the distance from the e-CP to the t-CP, and TND is the distance from the e-CP to the end point of the time series. These can be represented in formula (1) to formula (4).

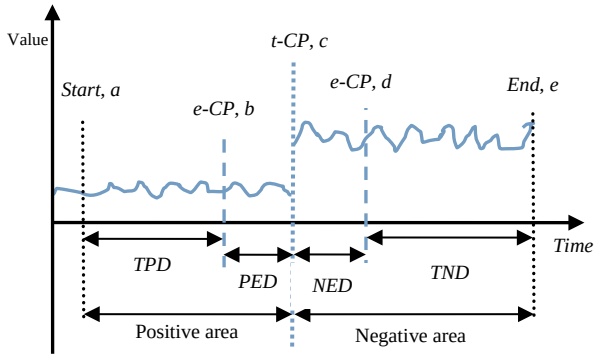

Fig. 1. The scheme of TPD, PED, TND, NED definitions in a model of change-point detection on one time-series. *a* represents start point position of the time series, *b* is the position of e-CP when it is on the left side of the t-CP, *d* is when e-CP is on the right side of the t-CP, *c* is the position of t-CP, and *e* represents the end point of the time series.

$$TPD = e\text{-}CP - Start = b - a \quad (1)$$

$$PED = t\text{-}CP - e\text{-}CP = c - b \quad (2)$$

$$TND = End - e\text{-}CP = e - d \quad (3)$$

$$NED = e\text{-}CP - t\text{-}CP = d - c \quad (4)$$

Where, *a* represents the position of the start point of the time series, *b* is the position of e-CP when it is on the left side of the t-CP, *d* is when it is on the right side of the t-CP. *c* is the position of t-CP, and *e* represents the position of the end point of the time series.

To measure the performance of a CPD model on multiple time series of varying lengths and t-CPs located at different positions, a normalized measurement metrics is designed. The normalized TPD, FND, TND and FPD can be represented as true positive distant rate (TPDR), positive error distance rate (PEDR), true negative distance rate (TNDR) and negative error distance rate (NEDR). These values can be calculated by formula (5) to formula (8). Basically, the distance between the start point and the t-CP and the distance from the t-CP to end point of each tested time series are both normalized to 1, and the normalized t-CP position for each time series will match to the same point. TPDR, PEDR, TNDR and NEDR can be interpreted as the positive weighted true distance (WPTD, weighted true distance in positive area), weighted positive error distance (PWED), weighted negative true distance (NWTD) and weighted negative error distance (NWED).

$$TPDR(WTPD) = \frac{TPD}{TPD+PED} \quad (5)$$

$$PEDR(WPED) = \frac{PED}{TPD+PED} \quad (6)$$

$$TNDR(WTND) = \frac{TND}{TND+NED} \quad (7)$$

$$NEDR(WNED) = \frac{NED}{TND+NED} \quad (8)$$

Suppose there are *N* time series, each contains a change point to be tested with a CPD model. For time series *i* (0 <*i*<*N*), its WTPD, WPED, WNTD and WNED can be illustrated as Fig. 2.

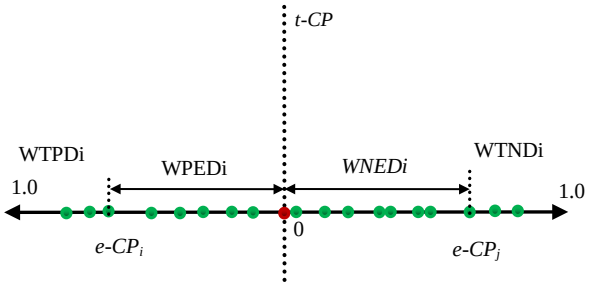

Fig. 2. The scheme of the proposed *WED* method for evaluating the search efficiency of a change-point detection model.

For evaluation of the overall performance of a CPD on multiple time series, both WPED and WNED for each time series *i* can be referred to as WED, shown in formula (9).

$$WED_i = \begin{cases} WPED_i & e\text{-}CP\ on\ positive\ area \\ WNED_i & e\text{-}CP\ on\ negative\ area \end{cases} \quad (9)$$

A mean weighted error distance (MWED) is defined as:

$$MWED = \frac{\sum_{i=1}^{m} WPED_i + \sum_{j=1}^{l} WNED_j}{m+l} \quad (10)$$

Where *m* and *l* refer to the number of the *e-CPs* located before and after the target *t-CP* (positive area and negative area) respectively. In most of the CPD models, when the search algorithm reaches the start or end point of the time series, if no change point is found, then the e-CP can be set as either the start or end point. The sum of *m* and *l* will be equal to *N* (number of time series to be tested by the CPD). The formula (10) can be simplified to:

$$MWED = \frac{\sum_{i=1}^{N} WED_i}{N} \quad (11)$$

Corresponding to a MWED, 1-MWED can be used as a measure of the overall performance for a CPD and be referred to as mean weighted true distance (MWTD).

To illustrate the positions of e-CPs when a CPD works on multiple time series, frequencies of WED values from the multiple tests can be calculated, and a histogram of WED showing the distribution of the normalized e-CP positions can be used to compare the performance between the CPD models. The frequencies of each WED value can be calculated as:

$$Freq(k) = Num(WED_k) * WED_k / N \quad (12)$$

N is the number of time series tested by the CPD, 0<*k*<*K*, *K* is the number of unique WED values of the tests on *N* time series (*K*<=*N*).

III. EXPERIMENTS AND RESULTS

We applied the proposed WED method and other existing measurements including hit rate, accuracy and computing time, to evaluate the performance of different CPD models including KS, SSA and T models. Hit rate is the probability of the e-CPs located at the target t-CP position, that is the ratio of the e-CPs occurred at the t-CP location in all N tested time series. Computing time is the actual time needed for the CPD to run the set of tests. For a CPD that tests single time series, accuracy is defined as 1-MAE [13], which can be calculated as:

$$1 - \text{MAE} = \left(1 - \frac{ED}{Length\ of\ time\ series}\right) * 100\%$$
$$= \left(1 - \frac{ED}{End - Start}\right) * 100\% \quad (13)$$

Where ED is the absolute distance between the e-CP and the t-CP. ED=PED, when e-CP is on the left side of t-CP, and ED=NED when e-CP is on the right side of t-CP. For multiple tests on N time series, average accuracy of the N tests is reported in this study, and is represented as "accuracy".

For experiments, multiple simulation datasets were created that included time series with different lengths, variances and t-CP positions. Each CPD model, including KS, SSA and T, was tested using the same datasets. Each time series included in a dataset that was used for testing the CPD had one t-CP. The positive area (start point to t-CP) of each time series $X_{il} = \{x_1, \ldots, x_m\}$ was composed of normal random numbers $N(\mu=0, \sigma=1)$ of size m (m time points included in the positive area). The negative area $X_{ir} = \{x_{m+1}, \ldots, x_N\}$ (from t-CP to end point) was simulated by adding a constant variance, V, into the normal random numbers $N(\mu=0, \sigma=1)$ of size L-m (L-m time points in the negative area), where L is the length of the time series.

Here we first present the result from a simulated dataset called Dataset1. Dataset1 included 1000 time series to be tested. Each time series was simulated with a random length between 64 and 512 time points. A fixed variance of 1.5 (V=1.5) was also used to create the values of the time series, following the procedure described in the paragraph above. The t-CP for each time series was set to a different position randomly. The experiment performed with this dataset is named Exp1.

The test results from Exp1, including hit rate, accuracy and computing time from different CPD modes, are listed in Table I. The results can help us to have an overall view of the performance of the CPD models. From these results, we can see KS produced the best result based on any of the measurements (hit rate, accuracy and computing time) used in this study. SSA produced higher accuracy which is based on the absolute error distance from the t-CP, but lower MWED than T model. In this case, the preference between the SSA or T models would depend on the application. If the absolute error matters, T method would be chosen while SSA would be a better choice if the relative error matters.

The advantage of the WED metrics is that it normalizes measures to enable the comparison of CPDs performance that work on time series with different lengths and with t-CP located at different positions. A histogram can be drawn to see the distribution of the WED across all the times series tested. This cannot be achieved with other metrics mentioned earlier in the paper. Fig. 3 shows the WED distribution with normalized e-CP frequencies (frequencies of each WED) calculated from formula (12) based on the result from Exp1. From the figure, we can see the frequencies of lower WED (both positive and negative, or WPED and WNED) values are higher than those higher WED values in KS model. Most of the WED values produced by KS model on the multiple time series included in Dataset1 were quite low. This means that most of the e-CPs from KS were distributed close to t-CP. The histogram drawn with the result from SSA and T models were more spreaded to the side of the x-axis. Quite a few higher WED values appeared with high frequencies from the result of T model. The visualized results indicated higher overall performance from KS across multiple time series with variety of lengths and t-CP positions.

TABLE I. CPD RESULT FROM KS, SSA AND T MODELS, USING DATASET1.

Methods \ Items	1-MWED (MWTD)	Hit rate	Accuracy	Computing time
KS	.9518	.1060	.9977	.0022
SSA	.8422	.0610	.9287	.0400
T	.3100	.0360	.9670	.0165

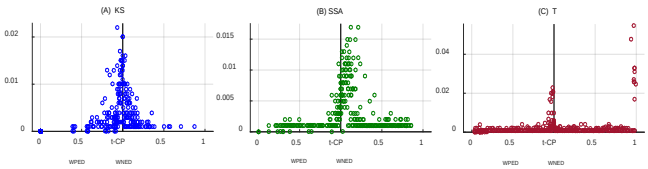

Fig. 3. Distrribution of WED from different CPD models. On x-axis, the left area represent the positive area and right negative area (normalized). y-axis is the frequency of each WED value from the multiple tests (1000 time series).

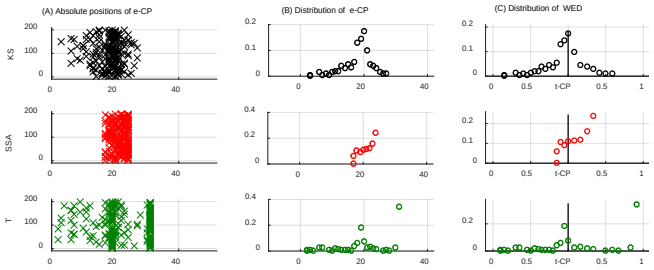

Fig. 4. Distrribution of e-CP or WED from different CPD models, using Dataset2 (N=200, L=32, V=1.5, t-CP=20). (a) Absolute positions of e-CP, x-axis is the positions of each time series which has length of 32, y-axis represent the 200 tests on the total number of time series. (b) Probability distribution of e-CP, x-axis is the positions of each time series, y-axis is the probability/frequency of the e-CP hit the position. (c) WED distribution, the normalized e-CP distribution. x-axis is the normalized the position of each time series (Fig. 2), y-axis is the probability of the e-CP hit the position.

When a CPD is used to detect the change points from multiple time series with same lengths and fixed location of actual change points, an e-CP distribution histogram can be produced without WED metrics. The probability distribution histogram should be the same shape as the histogram produced with WED metrics. Fig. 4 illustrates the result from the CPDs tested on a dataset called Dataset2 that included 200 time series with the same length. We call this set of experiments Exp2. The set up for Dataset2 is L=32, V=1.5 and t-CP=20. From Exp2, We can see that using WED metrics for e-CP distribution analysis presented the same result as the

earlier methods. However, the earlier methods are restricted to visualize the e-CP result only when the CPD worked on time series with the same length and t-CPs were located at the same position.

For this study, experiments were also performed with datasets created with fixed variances (e.g. V=0.5 or V=2) for each time series included in the same subset of dataset. Dataset3, for example, was created as 4 subsets (A, B, C, D), which included time series with variances V=0.7, V=1.6, V=2.5 or V=3.4 respectively. Each subset A, B, C or D, included 200 time series with a t-CP randomly created at different positions. The corresponding set of experiments on Dataset3 refers to Exp3. Fig. 5. shows the WED distribution from the CPD results working on subsets from Dataset3.

In general, the CPDs performed well on the time series with higher variances. Fig. 5. shows the normalized e-CP distribution based on the result from Exp3, with the WED values produced from KS, SSA and T models. The visualization based on the WED metrics made the results from different models worked on multiple various time series comparable.

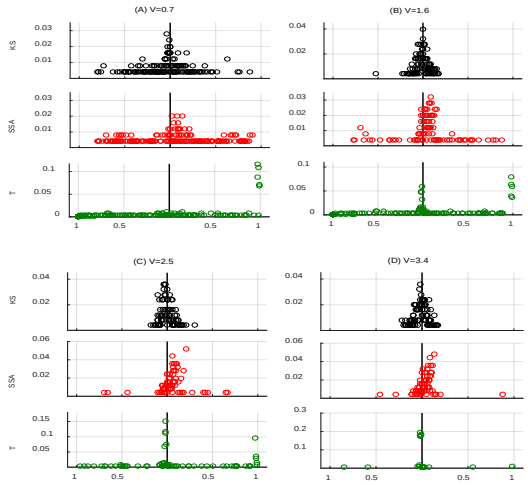

Fig. 5. The WED distribution from the CPD results on KS, SSA, and T methods, using different datasets (A,B,C,D) that each included 1000 time series in length 128. The t-CP was randomly set for each time series. A) each time seariers with V=0.7, (B) each time seariers with V=1.6, (C) each time seariers with V=2.5, and (D) each time seariers with V=3.4.

The overall evaluation of the CPDs tested using Dataset3, time series with different variances, can be done with the MWED values. The MWED results along with the hit rate and accuracy produced from each CPD model are shown in Table II. The ranks of the three models are the same based on the three metrics, which is the consistency that is expected in most of the real-world applications.

TABLE II. CPD RESULTS FROM KS, SSA AND T MODELS, USING DATASET3.

Items / Methods	1-MWED	Hit rate	Accuracy
KS	.9151	.1096	.9883
SSA	.8507	.0562	.9668
T	.6221	.0246	.9023

IV. CONCLUSION AND DISCUSSION

In this study, a new WED method that can be used for CPD performance evaluation is proposed. In this method, both positive and negative error distances from the CPDs are weighted or normalized for creating a WED metrics. As opposed to previous methods, WED values produced from CPDs using the new WED method allows comparison between the models, when CPD is used across multiple time series of different lengths and t-CP are located at different locations. The method was applied on evaluation of the CPDs utilizing KS, SSA and T methods. The results of the study showed its ability to compare the results from the CPD models working with multiple time series. The WED metrics offers a new way for evaluating CPD performance. It allows better visualization of distribution of the e-CPs when the CPD models work on multiple time series with different parameter values. Along with other evaluation methods, for example computational cost and hit rate, it can offer an overall measure and give better advice for users as to what CPD models to use based on the application. While technology of cloud and Interet of Things is growing fast, more biosignal data are collected for disease diagnosis and health care [20, 21]. It is important to develop CPD models and have them evaluated for proper applications, which will help improve health serivices.

REFERENCES

[1] R. J. Bolton and D. J. Hand, “Statistical fraud detection: A review,” Statistical Science, vol. 17, no. 3, pp. 235-249, 2002.

[2] K. Yamanishi, J. Takeuchi, G. Williams and P. Milne, “On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms,” Data Mining and Knowledge Discovery, vol. 8, pp. 275-300, 2004.

[3] U. Murad and G. Pinkas, “Unsupervised profiling for identifying superimposed fraud,” in Principles of Data Mining and Knowledge Discovery, Springer, 1999251-261.

[4] J. Reeves, J. Chen, X. L. Wang, R. Lund and Q. Q. Lu, “A review and comparison of changepoint detection techniques for climate data,” Journal of Applied Meteorology and Climatology, vol. 46, pp. 900-915, 2007.

[5] Y. Wang, C. Wu, Z. Ji, B. Wang and Y. Liang, “Non-parametric change-point method for differential gene expression detection,” PloS one, vol. 6, p. e20060, 2011.

[6] V. Pillai, D. Kalmbach and J. Ciesla, “A meta-analysis of electroencephalographic sleep in depression: evidence for genetic biomarkers,” Biological Psychiatry, vol. 70, no. 10, pp. 912-919, 2011.

[7] A. de Luna, I. Cygankiewicz, A. Baranchuk, M. Fiol, Y. Birnbaum, K. Nikus, D. Goldwasser, J. Garcia-Niebla, S. Sclarovsky, H. Wellens and G. Breithardt, “Prinzmetal angina: ECG changes and clinical considerations: a consensus paper,” Annals of Noninvasive Electrocardiology, vol. 19, no. 5, pp. 442-453. doi: 10.1111/anec.12194., 2014.

[8] J. P. Qi, Q. Zhang, Y. Zhu and J. Qi, “A novel method for fast Change-Point detection on simulated time series and electrocardiogram data,” Plos One, vol. 9, p. e93365, 2014.

[9] V. Moskvina and A. A. Zhigljavsky, “Application of singular-spectrum analysis to change-point detection in time series,” Cardiff, UK, 2001.

[10] J. P. Qi, J. Qi and Q. Zhang, “A Fast Framework for Abrupt Change Detection Based on Binary Search Trees and Kolmogorov Statistic (BSTKS),",” Computational Intelligence and Neuroscience, vol. 2016, p. 10.1155/2016/8343187, 2016.

[11] B. Darkhovski, “Nonparametric Methods in Change-Point Problems: A General Approach and Some Concrete Algorithms,” Lecture Notes-Monograph Series, vol. 23, pp. 99-107, 1994.

[12] D. Cook, Krishnan and NC, Activity Learning: Discovering, Recognizing, and Predicting Human Behavior from Sensor Data, Wiley, 2015.

[13] S. Aminikhanghahi and D. J. Cook, “A Survey of Methods for Time Series Change Point Detection,” Knowledge and Information Systems, vol. 51, no. 2, p. 339–367, 2017 .

[14] M. S. Pepe, “Receiver Operating Characteristic Methodology,” Journal of the American Statistical Association, vol. 95, pp. 308-311, 2000.

[15] J. A. Hanley, “Receiver operating characteristic (ROC) methodology: the state of the art,” Critical Reviews in Diagnostic Imaging, vol. 29, pp. 307-335, 1989.

[16] M. Grzybowski and J. G. Younger, “Statistical methodology: III. Receiver operating characteristic (ROC) curves,” Academic Emergency Medicine, Official Journal of the Society for Academic Emergency Medicine, vol. 4, p. 818, 1997.

[17] V. Moskvina and A. Zhigljavsky, “An Algorithm Based on Singular Spectrum Analysis for Change-Point Detection,” Communications in Statistics - Simulation and Computation, vol. 32, pp. 319-352, 2003.

[18] Z. Yang, K. T. Fang and S. Kotz, “On the Student's t-distribution and the t-statistic,” Journal of Multivariate Analysis, vol. 98, p. 1293–1304, 2007.

[19] J.-P. Qi, Q. Gu, Y. Zhu and P. Zhang, “A KST framework for correlation network construction from time series signals,” in Proceedings of the Ninth International Conference on Graphic and Image Processing, Qing Dao, 2017. R. J. Bolton and D. J. Hand, "Statistical fraud detection: A review," *Statistical Science*, vol. 17, no. 3, pp. 235-249, 2002.

[20] M. Haghi, K. Thurow, I. Habil and R. Stoll, "Wearable Devices in Medical Internet of Things: Scientific Research and Commercially Available Devices," Healthcare Informatics Research, vol. 23, no. 1, pp. 4-15, 2017.

[21] M. Elgendi, "Less Is More in Biosignal Analysis: Compressed Data Could Open the Door to Faster and Better Diagnosis," Diseases, vol. 6, no. 1, p. pii: E18. doi: 10.3390/diseases6010018, 2018.