On application of distance-like algorithms to event detection from non-stationary time series[1]

Tomasz Pełech-Pilichowski

AGH-University of Science and Technology, Krakow, Poland

tomek@agh.edu.pl

[1] This work was supported by the European Regional Development Fund, Grant no. UDA-POIG.01.03.01-12-171/08/00

***Abstract*– In the paper, event detection from time series with distance-based detection algorithms is described. Classical distance measures and their applicability to non-stationary time series data processing are presented and reviewed. Conditions for event detection from diagnostic signals containing time-lagged events are investigated. Two sample distance-like detectors dedicated to identification of original changes in analyzed signals are introduced.**

***Keywords*– time series, event detection, distance measures**

I. INTRODUCTION

In recent years, both the availability of computer systems and information technology development create new possibilities for capturing, collecting and sharing large datasets. Such process creates opportunities and – on the other hand – the need to improve an automation of information selection, which is very important for decision making process.

Large time series datasets received from multiple technical devices (usually networked), are processed (online/offline) and they can be a valuable source of implicit information. Moreover, for advanced control systems, due to the real-time regime, the availability, selection and processing of diagnostic signals are vital for both the basic data processing (usually performed in real-time) and application of dedicated numerical procedures (control optimisation, predictive control etc.). Such procedures are based on accurate statistical parameters estimation and require implementation of efficient algorithms for detecting changes in time series statistical properties. For this purpose, one may use classical methods of signal analysis (statistical and frequency ones [2]), data mining algorithms [11], computational intelligence procedures [9] or multivariate time series analyses [27]. Data mining and event-detection algorithms are often based on examination of similarity between objects [11]. Such measures are able to compare processed subseries and identify occurring differences as short- and long term changes.

The aim of this paper is to present a novel approach to event detection from time series based on distance-like change detectors. Requirements for efficient analysis and event detection from non-stationary datasets, including diagnostic signals containing similar time-lagged events are studied. Common approach to time series monitoring based on distance-based algorithms is reviewed, standard metrics are briefly described. Two sample distance-like event detection algorithms aimed at untypical change detection are introduced.

II. TIME SERIES COMPUTER MONITORING

Time series monitoring is essential for commercial and individual use, as well as scientific research aimed at implementation of numerical procedures for computer control and supervision. The aim is to catch and select relevant information with decomposition into slow and fast-changed components. Such decomposition may be achieved with low-pass filtering, i.a. time series smoothing method [29], wavelet transform [29], least-squares approximation [7] and signal models identification. Computer selected information may be further used for expert system-based analyses or quantitative processing. Dedicated algorithms are exploited in many fields as technical diagnostics [4], medicine [15], pattern recognition [21], defence etc.

Obtained results of time series processing are vital for prediction and event detection, wherein *the event* can be viewed as unusual system behaviour which causes the short-, medium- or long-term changes of statistical or frequency properties in processed subseries, outliers or short sequences of samples forming patterns [21], [12]. Such changes may be preceded by symptoms revealed as short-term changes of a specific configuration; therefore, event detection is an interesting area from signal analysis and artificial intelligence point of view.

Considering two time series (training a testing ones), *the symptoms of event* may be defined as significant difference (assuming to specific criteria, for example statistical ones) between two time series in a fixed time interval [16]. Most of events are visible (explicit) in processed data as rapid changes in signal level or as atypical values. Nevertheless, there are hard-detectable (implicit) events which may precede long-term changes of statistical properties in one of concurrent processed subseries.

Digital processing of available diagnostic signals set requires an application of algorithms for detecting changes in statistical properties of time series (for example, abrupt changes of mean value, short-term signal changes etc). Dedicated procedures should allow indicating unusual patterns, novelties, anomalies or outliers in analyzed datasets (for example – considering signals received from technical devices – alarm notifications, faults; considering financial time series – long-term changes indicators, uncertainly in the markets)

Event detection algorithms have applicability in many areas, such as detection of damages (fault detection) [5], pattern recognition [21], analysis of computer network traffic

[18], communication systems [10], prediction [20], online services and e-commerce, statistical process control (SPC) [9], [19], marketing [1] etc.

To obtain efficient detection with classical methods (statistical and frequency ones) long datasets are required. This requirement is often unacceptable; therefore, researchers have investigated approaches for both accurate and fast detection process. Described and implemented advanced algorithms usually are based on computational intelligence and artificial intelligence paradigms (machine learning [8], artificial neural networks [17], artificial immune systems [4], [5], expert systems [28], fuzzy logic [28]) or multivariate time series analysis [27]. Many well-defined and efficient detection procedures are based on data mining and knowledge discovery techniques [11], [14], [16], [19], [25] which in many implementations (e.g. nearest-neighbour method, cluster analysis, multidimensional scaling method [11]) explore similarity (or dissimilarity) methods are employed to reveal differences between objects, where a selection (definition) of similarity function is crucial.

III. Distance Measures

To study the degree of similarity (dissimilarity) the distance or metric term is used (a function determining the distance between objects). A metric d is a measure satisfying the four conditions for each i, j, k [11], [27]:

1. $d(i, j) \geq 0$, for all i and j;
2. $d(i, j) = 0$, if and only if $i = j$;
3. $d(i, j) = d(j, i)$, for all i and j (symmetry);
4. $d(i, j) \leq d(i, k) + d(k, j)$, for all i, j and k (triangle inequality).

To analyze similarity of time series of fixed length (in a constant moving window), in a quantitative (numerical) view, one may use distance measures with one-dimension-conversion – which allows to eliminate an impact of different dimensions of analyzed object, it also simplifies a comparison of obtained computation results. Low distance values indicate a high similarity between objects, while high values – dissimilarity of processed series.

The most common measure of similarity between objects y and x (e.g. time series of the length p) is defined as the Euclidean distance [11],[27],[25]:

$$d_E(y,x) = \sqrt{\left(\sum_{k=1}^{p}(y_k - x_k)^2 \frac{1}{p}\right)} = \sqrt{\frac{(y-x)'(y-x)}{p}}$$

where p – number of samples of each object (the size of the feature space).

Euclidean metric is considered as vital basis for classification [3]. It is a generalization of the Minkowski metric [11] (also called m-norm):

$$d_M(y,x) = \left(\sum_{i=1}^{p}|y_i - x_i|^m / p\right)^{1/m}$$

Considering Minkowski metric, a comparable values of objects y and x (of the same order of magnitude) are assumed, which is achieved by standardization of processed subseries.

Besides Minkowski-based methods (e.g. City, Mahalanobis [11], [25]), there are a number of distance measures proposed by researchers, like Spearman Rang Coefficient, Kendall Tau Rank Correlation Coefficient [26] etc.

Measures aimed at finding dissimilarity between training series and test sets were proposed by E.Keogh (e.g. Dynamic Time Warping (DTW) [25], Compression Dissimilarity Method (CDM) [13], [16]). Notice, that such measures, compared to Euclidean, are valuable only for specific assumptions (conditions).

In the context of time-lagged events (changes, anomalies, deviations) in a processed non-stationary time series, an application of most widely exploited similarity measures (i.a. Minkowski-based) is not sufficient to identify *the real difference* between processed subseries, usually during analyses performed in a moving window which may cover different events for successive iterations. It results in decreasing in the reliability of event detection task, thus, such measures are often unsuitable as robust event detectors.

IV. Event Detection With Distance-Like Algorithms

A. Distance-like event detectors

Many methods of time-series analysis (detection and prediction algorithms) are suitable for stationary time series data, therefore, such procedures usually don't provide acceptable results when applied to hard-predictable non-stationary datasets.

Volatility of variance and mean value requires the analyses performed in a fixed moving window of relatively short width and based on subseries processing with distance-based procedures. In this context, it can be assumed that the similarity of two series consisting of relatively small number of samples may be viewed as the similarity of their short-range characteristics (statistical properties) and therefore as the detector of changes in time series.

In the proposed approach, event detection from time series is viewed as recognition of *real differences* between two subseries, i.e. events presented only in one processed signal (subseries) of fixed length. Such approach, when compared to widely and common used time series similarity methods [11], allows avoiding false alarms and undetected events, for example in the following cases:

- both subseries contain similar time-lagged events but only one event is covered by the analysis window;
- signals contain events for the same samples but with different sign (reverse events);
- subseries contain events for the same samples but with different attributes and configurations.

Notice, that accurate detection with distance-based methods should focus on the analysis both event presence in time series and similarity of events (and their attributes).

To achieve accurate detection and obtain unified (standardized) datasets, input data should be pre-processed. Such operation may be accomplished – depending on statistical and frequency data properties – with the following steps:

1. signal centering (substracting the average or detrending),
2. unifying – dividing by the reference value (external, standard deviation, absolute maximum etc.).

Such unified diagnostic signals may be further converted with transformations appropriate to a particular detection problem or required input data properties.

B. Processing of time series containing time-lagged events

The most important information concerning unusual series behaviour includes deviation amplitude, duration and delayed events in both processed series. Considering non-stationary time series (e.g. financial ones) which often contain many non-random components, due to the variability of delayed events the analysis reliability may be significantly reduced (notice, that statistical measures, such correlation coefficient assume constant delays). Therefore, efficient detection with distance-based method can be obtained with applying *the tolerance* during calculations or with the use of measures based on amplitude spectra.

Based on recent work [24], to avoid the impact of time delay between events, for each sample t a number (denoted as L_{tol} – called *the tolerance* – permissible delay between analyzed signals) of distance measures d_n is computed ($n = t - L_{tol} + 1,...,t$). The final distance measure value between two subseries for sample t is taken as the lowest d_n value obtained.

C. Sample distance-like detection algorithms

To satisfy restrictions resulting from non-stationary time series processing, two sample algorithms are presented [24], dedicated to catch specific (original) deviations in two processed series. Notice, that in this case "distance measure" term is used instead of "distance" or "metric" because all requirements related to formal criteria of metric definition are not satisfied (for example, symmetry condition – see Section III) .

1) Measure of unified patterns similarity

The proposed detection method (denoted as U) [6], [24] is designed to identify *unique* changes recognized in two processed subseries of the length N (constant moving window width N is assumed) as subsequences of deviations of the same sign exceeding an arbitrary fixed threshold ρ_U.

The aim is to detect unique subsequences of different length (1,2,….,N) in one diagnostic signal with no reference in the second analyzed one (events are not similar).

The main parameter of measure U is the threshold ρ_U which value may be fixed as multiplicity of variance (computed in a moving window) or any value related to significant change properties. ρ_U may be also adapted depending on statistical properties of signals.

The detection process includes the following steps:

a) Analysis of sequences of changes of different length (1,2,…N) in the both processed subseries (x and y)
b) K_k calculation – as the maximum length of detected subsequences of deviations
c) L_{kx} and L_{ky} calculation – as the number of subsequences detected in x and y (where k =1,2,…,K_k)
d) w_{pzg} calculation – percentage of coincidence sequences in two subseries as follows (for similar detected sequences in both subseries, w_{pzg} value is close to 1):

$$w_{pzg} = \left(1 \Big/ \left(\sum_{k=1}^{Kk} k \cdot \max(L_{kx}, L_{ky})\right)\right) \cdot \sum_{k=1}^{Kk} k \cdot \min(L_{kx}, L_{ky})$$

e) Finally, the distance measure d_U calculation with the following formula:

$$d_U = 1 - w_{pzg}$$

According to test analysis performed on sample non-stationary data [23], method U has applicability for single, concurrent patterns. It is also vital for time series short- and long-term event detection [24].

2) Event-driven similarity

The second presented distance-like similarity method (denoted as Z) [22], [6] is aimed at synchronous processing of two signals x and y of fixed length (analyzed in a moving window) with comparison of changes exceeding a fixed threshold ρ_{Zd}.

In particular, the detection process includes the following steps:

a) For x and y, mean values of positive deviations (x_{pm}, y_{pm}) exceeding ρ_{Zd} are calculated (notice, subseries mean value close to zero is assumed; it which may require data differentiation)
b) For x and y, mean values of negative deviations (x_{nm}, y_{nm}) exceeding ρ_{Zd} are calculated
c) The distance measure d_Z is calculated employing the following formula:

$$d_Z = \sqrt{(x_{pm} - y_{pm})^2 + (x_{nm} - y_{nm})^2}$$

Referring to preliminary analysis of detection effectiveness [23], [22], method Z is valuable for identification of concurrent patterns in both time series and as a large original change (deviation presented only in one processed subseries) detector.

V. CONCLUSIONS

Time series processing aimed at accurate monitoring and efficient event detection requires the use of dedicated algorithms, including distance-like ones. It is relevant approach in many areas of diagnostic signal processing, especially for real time systems operations. Therefore, research focused on novel event detectors defining (designing) and testing is significant.

In this paper, constraints resulting from implementation of classical distance metrics have been emphasized. Nevertheless, general approach to robust event detection based on distance-like procedures has been introduced. It implies that (1) data pre-processing significantly affects the obtained analysis results (in particular, data centering and unifying); (2) to avoid unreliable detection results when processing time series data containing time-lagged events, distance measures should be calculated with a tolerance and

(3) considering non-stationary diagnostic signals consist of many random and non-random components, procedures capable of detecting original changes should be dedicated, i.e. designed and tested for specific (original) change identification (in the paper, two sample dedicated algorithms have been introduced).

Further research will focus on designing algorithms aimed at detection of another untypical changes presented in processed diagnostic signals, including different configurations of deviations and patterns. Moreover, testing procedures of proposed algorithms performed on real data received from MES/SCADA systems are planned.

REFERENCES

[1] Aggarwal C.C., Yu P.S.: Outlier Detection for High Dimensional Data. ACM International Conference on Management of Data, Vol. 30, Issue 2, 2002

[2] Brockwell P.J., Davis R.A.: Time Series: Theory and Methods. Springer Series in Statistics, 1991

[3] Ciufudean C., Larionescu A., Filote C., Mahalu G.: Modeling Immunology Mechanisms with Discrete Event Systems. European Conference on Intelligent Systems and Technologies, 2004

[4] Dasgupta D., KrishnaKumar K., Wong D., Berry M.: Negative Selection Algorithm for Aircraft Fault Detection. International Conference on Artificial Immune Systems, 2004

[5] Dasgupta D., Nino F.: A Comparison of Negative and Positive Selection Algorithms in Novel Pattern Detection. IEEE International Conference on Systems, Man and Cybernetics (SMC), Vol. 1, 2000

[6] Duda J.T., Pełech-Pilichowski T.: Miary podobieństwa szeregów czasowych w detekcji zdarzeń. [In:] Systemy wykrywające, analizujące i tolerujące usterki / red. Kowalczuk Z., PWNT, 2009

[7] Fernandez V.: Does Domestic Cooperation Lead to Business-Cycle Convergence and Financial Linkages? The Quarterly Review of Economics and Finance, Vol. 46, Elsevier, 2006

[8] Glickman M., Balthrop J., Forrest S.: A Machine Learning Evaluation of an Artificial Immune System. Evolutionary Computation Journal, Vol. 13, No 2, 2005

[9] Guh R., Zorriassatine F., Tannock J.D.T., O'Brien C.: On-line Control Chart Pattern Detection and Discrimination - a Neural Network Approach. Artificial Intelligence in Engineering, Vol. 13, Issue 4, Elsevier, 1999

[10] Grzech A.: Anomaly Detection in Distributed Computer Communication Systems. Cybernetics and Systems, vol. 37, nr 6, 2006

[11] Hand D., Mannila H., Smyth P.: Principles of Data Mining. MIT Press, 2001

[12] Hajnicz E.: Time Structures. Formal Description and Algorithmic Representation. Lecture Notes in Computer Science , 1996, Vol. 1047, p.4

[13] Keogh E., Lonardi S., Ratanamahatana C.: Towards Parameter-Free Data Mining. ACM International Conference on Knowledge Discovery and Data Mining, 2004

[14] Lazarevic A., Kumar V.: Feature Bagging for Outlier Detection. ACM International Conference on Knowledge Discovery in Data Mining, 2005

[15] Lin J., Keogh E., Fu A., Van Herle H.: Approximations to Magic: Finding Unusual Medical Time Series. IEEE Symposium on Computer-Based Medical Systems, IEEE Computer Society, 2005

[16] Mahoney M.V., Chan P.K.: Learning Rules for Time Series Anomaly Detection. Technical Report CS-2005-04, Florida Institute of Technology, 2004

[17] Mäyrä O., Ahola T., Leiviskä K.: Time Delay Estimation and Variable Grouping Using Genetic Algorithms. Control Engineering Laboratory, Report A No. 32, University of Oulu, 2006

[18] Muthukrishnan S., Shah R., Vitter J.S.: Mining Deviants in Time Series Data Streams. International Conference on Scientific and Statistical Database Management, IEEE Computer Society, 2004

[19] Paavola M., Ruusunen M., Pirttimaa M.: Some Change Detection and Time-series Forecasting Algorithms for an Electronics Manufacturing Process. Control Engineering Laboratory, Report A No. 26, University of Oulu, 2005

[20] Palit A.K., Popovic D.: Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications. Springer, 2005

[21] Papadimitriou S., Sun J., Faloutsos C.: Streaming Pattern Discovery in Multiple Time-Series. International Conference on Very Large Data Bases, 2005

[22] Pełech-Pilichowski T.: A hybrid algorithm of event detection in diagnostic signals received from intelligent building devices, Unpublished paper, under reviewing process, Intelligent Building International, 2010

[23] Pełech-Pilichowski T.: Adaptacyjne algorytmy detekcji zdarzeń w szeregach czasowych, Ph.D. Thesis, AGH-UST, 2009

[24] Pełech-Pilichowski T., Duda J.T.: A two-level algorithm of time series change detection based on a unique changes similarity method, International Multiconference on Computer Science and Information Technology Proceedings, Wisła, Poland, 2010

[25] Shyu M., Chen S., Sarinnapakorn K., Chang L.: A Novel Anomaly Detection Scheme Based on Principal Component Classifier. IEEE International Conference on Data Mining, 2003

[26] Taylor J.M.G.: Kendall's and Spearman's Correlation Coefficients in the Presence of a Blocking Variable. Biometrics. Vol. 43, No. 2, 1987

[27] Yang K., Shahabi C.: A PCA-Based Similarity Measure for Multivariate Time Series. ACM International Workshop on Multimedia Databases, 2004

[28] Zhang Y., Akkaladevi S., Vachtsevanos G., Lin T. Y.: Granular Neural Web Agents for Stock Prediction. Soft Computing – A Fusion of Foundations, Methodologies and Applications, Vol. 6, No. 5, Springer, 2002

[29] Zieliński T.: Reprezentacje sygnałów niestacjonarnych typu czas-częstotliwość i czas-skala. Wyd. AGH Kraków, 1994