A SURVEY OF RESEARCH ON ANOMALY DETECTION FOR TIME SERIES

HU-SHENG WU

Materiel Engineering College, Armed Police Force Engineering University, Xi'an 710086, China
E-MAIL: wuhusheng0421@163.com

Abstract:

Time series is an important class of temporal data objects and it can be easily obtained from scientific and financial applications, and anomaly detection for time series is becoming a hot research topic recently. This survey tries to provide a structured and comprehensive overview of the research on anomaly detection. In this paper, we have discussed the definition of anomaly and grouped existing techniques into different categories based on the underlying approach adopted by each technique. And for each category, we identify the advantages and disadvantages of the techniques in that category. Then, we provide a briefly discussion on the representative methods recently. Furthermore, we also point out some key issues about multivariate time series anomaly. Finally, some suggestions about anomaly detection are discussed and future research trends are also summarized, which is hopefully beneficial to the researchers of time series and other relative domains.

Keywords:

Time series; anomaly detection; big data; data mining; multivariate time series

1. Introduction

Date have proliferated over with the development of information technology. According to the latest research of University of Southern California in 2016, since 1980s, the scale of data has increased sharply and doubled every year or even a few months. All of the global data information has reached 295EB until 2007, 1.8ZB until 2011, and will reach more than 40ZB in 2020. Therefore, big data era has arrived. And how to abstract potential information from the plentiful data needs to be solved urgently [1]. A big part of these data is time series. The so-called time series is a sequence records according with the chronological order. Time series is so widely applied in many fields, such as military, economy and scientific observations that it has aroused great attention concern from researchers.

For normal time series data, although the number of abnormal data is very small, it does not mean that the abnormal data is not important. On the contrary, some important data may be hided behind these few abnormal data. In medical field, the abnormal heart beat should be detected so that doctors can find the disease in time by the ECG observation. In addition, the anomaly detection of time series can be widely used in the engine condition monitoring, network intrusion detection, anti-money laundering, network public opinion monitoring, credit card fraud, stock market analysis, improper tax behavior monitoring, major construction projects inspection, natural disaster analysis and so on. Obviously, the study of abnormal detection of time series has important theoretical value and practical significance.

2. Definition of time series anomaly detection

Time series widely exist in various large financial, medical, engineering and social science database. There are two features between time series and other data types. Firstly, the time attribute, the records of each variable must have time dimension and should be arranged in chronological order, and some date types like market basket data do not have such attributes. Secondly, the sequence attribute, the record values are a continuous one in a certain period of time and in certain laws. Time series can be divided into univariate time series and multivariate time series according to the number of variables, the specific definition is as follows:

Definition 1 Time Series A sequence records according with the chronological order $S=\{v_i(1),v_i(2),\cdots\cdots,v_i(t),\cdots\cdots,v_i(n)\}$ is defined as the time series. In which t (t=1,2,$\cdots$,n) is defined as time, i ($i=1,2,\cdots,m$) is defined as variable. $v_i(t)$ is defined as the record of the variable i on time t. When m=1, S is defined as univariate time series (UTS). When m>1, S is defined as multivariate time series (MTS).

At the same time, in order to make the anomaly detection of MTS meaningful, MTS needs to meet the three requirements as follows: (1) For the MTS data set, the variable dimension of each sequence is identical, and the variables which have the same meaning are one-to-one match. (2) For one MTS sample, the recording time of each

variable is the one to one correspondence and have the same time granularity. (3) MTS needs to be regularized.

To study time series anomaly detection, the first thing is the definition of the anomaly, that is, what is the exception. There is not a harmonized definition agreed by academia for exceptions. In statistics, the data points which do not obey the sequence distribution and far from other objects are defined as abnormal. In regression model, the data points which largely diverge from the designated model are defined as abnormal. Many researchers also proposed some similar terms according to their own understanding such as Anomaly, Surprise, Outlier, Novelty, Change Point, Devian and so on [2]. The highly accepted definition was proposed by Kawkins [3]: the abnormal was the data which was very different from others in the data sets and the causes of it was not due to random errors but different mechanisms. Therefore, most scholars provided the definition of abnormal in time series combining with actual applications on the basis of Kawkins' conclusion.

3. Method of time series anomaly detection

Many foreign scholars have joined in the studying on time series anomaly detection since Barnett published the first book Outliers in statistical data on anomaly detection in 1980s, such as M Breunig, E M Knorr, E Keogh, Portnoy, J Takeuchi, M Agyemang, M Markou, V Chandola and so on. The domestic research begins very late but developed rapidly. The relevant researches are carried by Tsinghua University, Xi'an Jiao Tong University, Tianjin University, Fudan University, Hong Kong University of Science and Technology, etc. Because of the research value and application prospect of time series anomaly detection, a large number of scholars have joined in its study. Some high-quality achievements on UTS were achieved on the famous international conference such as PAKDD、PKDD、SIGKDD、VLDB in the past ten years and published in the journals like IEEE TKDE 、 Neural Computation ， Computational Statistics and Data Analysis.

As an important sub-branch of data mining, anomaly detection is receiving more and more attention and research. Domestic and foreign scholars have proposed many methods of anomaly detection, which can be divided into five categories as follows: abnormal detection based on statistics, abnormal detection based on clustering, abnormal detection based on distance, abnormal detection based on density, etc [4].

3.1. Anomaly Detection Based on Statistical

Anomaly detection based on statistical is the earliest and most studied approach. It considers that the difference between the data and the statistical distribution or the model given is bigger than a particular value or range that is anomaly [5]. This method can be divided into two categories: distribution-based method and depth-based method. The former is to assign a distribution (such as normal distribution, Poisson distribution, etc.), and then the anomaly is found by the method of consistency checking. The vast majority tests of anomaly detection based on distribution are for a single attribute while a large number of abnormal data mining in reality is carried out in a multidimensional space. In addition, the actual distribution of the data set is often unknown, and it is difficult to estimate the data distribution in high-dimensional; The latter method considers that each object is a point in n-dimensional space, each point has one set depth, and anomaly possibly exists in the object which has lower depth. This process avoids the data distribution fitting problem of distribution method of distribution based method and has better efficiency on UTS detection. But it needs to calculate the convex closure of n-dimensional space and has higher calculation complexity, only suitable for low dimensional data like two dimensional or three, and has a low efficiency on large data sets with four or higher dimensions [6].

3.2. Anomaly Detection Based on Clustering

Anomaly detection based on clustering clusters data directly or indirectly using the existing clustering algorithms, such as DBSCAN, ROCK, K-Means, etc [7], which considers the exception of the class or the data that cannot be clustered as anomaly. A large number of existing research results can be used. But clustering analysis is different from anomaly detection, the former aims to find the category of clustering, and the latter aims to find abnormal data. Anomaly detection is only a "subsidiary products" of clustering. There is no special optimization in general clustering algorithm so that a low efficiency is caused. And in most cases, the definition and testing standard of anomaly is implicit, and cannot be clearly reflected in the process of clustering.

3.3. Anomaly Detection Based on Deviation

The anomaly detection based on deviation is mainly divided into three categories such as the sequence anomaly detection method, the OLAP data cube method and the prediction model method.

(1) Sequence anomaly detection method. Agrawal et al. proposed Sequential exception in 1996. It considered the data points with obvious deviation from adjacent

sequences as anomaly using a mechanism scan data sets. The computation complexity of this algorithm has a linear relationship with the size of data set. It has high efficiency. But its assumption of exception is too ideal, conceptually unsound, easy to leave out many useful anomaly data, and is ineffective in processing the real complex data.

(2) Sequence anomaly detection method. Sarasangi et al. used data cube technique to detect the anomaly area dealing with large-scale data set [8]. The unit values of a data cube are considered as anomaly if they are significantly different from the expected values of the statistical model. On this basis, the causes of abnormal can be found out by analyst drilling layer by layer according to data hierarchical structure. This method takes into account the changes in the metric values of all the dimensions of a cell and the exceptional cases which are hidden in the back of grouping operation of data cubes set. However, artificial detection is very difficult if there is large search space, especially multi-story concept hierarchy dimension [9].

(3) Sequence anomaly detection method. Many domestic and international scholars use Bayesian network, ARAM, neural network, support vector machine and other models to study the unknown endoplasmic relationship in time series data, and then establish the forecasting model and judge the abnormal through the deviation of the prediction value and the actual value [10]. This method has good effect on the data set with low variable dimension. But for high variable dimension, it is not appropriate for the abnormal detection of multi-variate time series because of the non-convergent training process.

3.4. Anomaly detection based on distance

The basic thought of anomaly detection based on distance is to calculate the distance between data points in data space by setting a distance function. It is regarded as abnormal when there is a large distance between a data object and others. Knorr et al. firstly proposed an anomaly detection method based on distance [11]. They consider o as DB（p, d） exception if the distance between object o and p objects at least is greater than d. And then, the concept of this distance is extended to k-neighbor distance. K-neighbor distance of each object is calculated and sorted from small to big by Ramaswamy et al, the objects which have largest distance are considered abnormal [12]. At present, there are several anomaly detection methods based on distance such as the algorithm based on index, the algorithm based on nested loop and the algorithm based on element [13]. The anomaly detection based on distance combines the ideas based on distribution, overcomes the primary disadvantages of anomaly detection based on distribution, is easier to realize and comprehend, and is widely studied and used. But it has its own drawbacks. Firstly, the complexity of the algorithm is relatively high, and it cannot take into account the data size of data sets and the scalability of dimension. The time complexity of the algorithm based on index and the algorithm based on nested loop can reach O (Mn2). The time complexity of the algorithm based on element is O (cm+n), in which m means the dimension, n represents the data object in data set, C represents the number of units; Secondly, Breunig et al. pointed out in Literature [9] that anomaly detection based on distance is flawed in processing the data sets with obvious internal density differences. It either considers the data in the area of sparse density to be abnormal or it cannot find some anomalous. This is mainly because of the anomaly detection based on distance considers all the viewpoints, its capability is not good in processing the data which contains a variety of distribution or are mixed with different densities subset. In addition, anomaly detection based on distance is very sensitive to parameters p and d, which requires the users having a certain expertise to set the reasonable parameters. So, its practical application is limited.

3.5. Anomaly Detection based Density

All the methods above have a common problem, that is, a global distance criterion is considered as the basis for anomaly detection. In fact, the anomaly is usually detected from the perspective of individual, that is, the anomaly point is far from its neighbor cluster. So it is not appropriate to use global distance. To solve this problem, Breunig and other scholars proposed anomaly detection algorithm based on density. Its basic idea is detecting anomaly by comparing the density of object and its neighbor. It introduces local outlier factor (LOF), considers the exception is not a two-value property but a measure. The LOF value higher, the data is more likely to be abnormal. Agyemang et al. proposed local sparsity coefficient (LSC） to detect anomaly and reduce the computational complexity by considering the largest distance between the nearest k objects as *k*-distance [14]. Furthermore, Papadimitriou and other scholars got multi-granularity deviation factor (MDEF） by comparing the number of data objects in r-neighbor and their mean values, took it the measure of abnormal[15]. This method does not need to calculate the density of data points, and the computational efficiency is higher than LOF. The idea based on density is closer to Hawkins' exception definition than the idea based on distance. So it can detect

the local anomaly, reduce the detection error which contains a variety of distribution or are mixed with different densities subset, have a higher detection precision [16]. However, there are some problems in the method based on density, its time complexity is still high, the detection results are sensitive to the selection of the parameters like outlier factor threshold and the parameters are difficult to determine.

4. Review of Typical Methods

One important feature of time series is time attribute, that is, there is a strict order between the sequence values. Time series belongs to ordered data. According to the characteristics of the time series, Chinese and foreign scholars had done a lot of research on anomaly detection. Now, we list some representative research achievements.

E. Keogh et al. signified UTS, retrieved the most significant sub time series by symbol [17]. It is simple and easy to implement, but how to describe the original data better is also a problem that is worth thinking.

S. Sadik et al. considered all the received data points as global context and the temporary closured data points as local context, presented automatic anomaly detection of data flow (A-ODDS) by detecting the departure between global context and local context in literature [18].

C. Shahabi et al. proposed the improved TSA-Tree algorithm and realized the abnormal pattern search through local maximum of wavelet coefficients in Literature [19]. But the abnormal pattern is based on the wavelet decomposition, so it may miss out some exceptional pattern in Literature.

V. Chandola used subspace monitoring converting MTS into UTS, and then used a WINCsvm method to implement anomaly detection in Literature [20]. It can simultaneously take into account the multiple and time sequence characteristics of MTS. But the calculation of the sliding window feature vector is the bottleneck of computing efficiency, especially when MTS has a large-scale size.

A method for anomaly detection based on density was presented by using the basic principle of Voronoi diagram in Literature [9]. It was applied to point anomaly of UTS and linear model. The algorithm complexity was reduced to $O(n\log n)$, but it had no concern with anomaly detection of MTS. It proposed a method based on distance and density, measured the abnormal degree of time series model based on GMBR (Grid Minimum Bounding Rectangle) of UTS by abnormal eigenvalues in Literature [21]. But its base theory research still need to be done in the future and bend in the direction of MTS anomaly detection.

In literature [22], they detected anomaly and fairly good practical results had been achieved by calculating the similarity between 2 MTS sub sequences using extended Frobenius norm, obtaining pattern set through k- means clustering, computing the exceptional support and frequency of each pattern. But how to enhance the efficiency of the algorithm in order to fit the MTS anomaly online recognition should be further discussed.

A two-stage MTS anomaly detection algorithm was proposed in paper [23]. It calculated the similarity between MTS samples using bounded coordinate system technology, realized anomaly detection using the method based on distance. The algorithm is divided into two stages: in first phase, they used K-means algorithm clustering and estimated the possibility of each cluster containing abnormal points. In second phase, they improved the efficiency of the algorithm by pruning rules on the basis of loop nesting algorithm. But it was still a anomaly detection method based on distance, which had a poor detection rate for the detection of time series data set which is mixed with different density subset.

A method of MTS anomaly detection based on KPCA was proposed in Literature [24]. It can be used in anomaly detection for different data by mapping the MTS data to the high dimensional feature space using kernel function invisibly, considering principal components orientation vector of data as feature expression using KPCA method. But its subsequence length and kernel function had a greater effect on the result. What's more, it cannot handle the relationships of variables well. So, this method should be further discussed.

In Literature [25], Naoya. T proposed a novel anomaly detection method for multivariate time series to capture relationships of variables and time-domain correlations simultaneously. The supposed framework in this study is a semi-supervised anomaly detection. The proposed method is based on feature extraction with sparse representation and relationship learning with dimensionality reduction. However, its computation complexity could be quite improved.

It can be seen that the abnormal detection of time series is not mature especially for MTS. That's mostly because of the following problems:

(1) Abnormal measurement problem. Because of the sparsity of MTS data, all the objects are likely to be abnormal in the sense of traditional distance. Therefore, how to define and measure the exception of multivariate time series is the primary problem.

(2) Complex data type. MTS type of data have the characteristics such as sequential, high dimension, complex correlation, noise interference, etc., making it more difficult to detect the anomaly.

(3) Curse of dimensionality. The curse of dimensionality is mainly indicating: the efficiency of the index structure is rapidly decreased with the increasing of the dimension, so that it is even not as well as the sequential scanning. In the high dimensional space, the nearest neighbor concept may become meaningless, because in many cases, the distance between a point and its nearest neighbor or furthest neighbor is almost equal [26]. It has brought much more difficulties for the rapid anomaly detection of time series.
(4) High computational complexity. Rapid anomaly detection algorithm largely depends on index structure and grid partitioning, its time complexity will increase and efficiency will decrease when the index structure loses efficacy in high dimension space or the number of grid partitioning increases exponentially with dimension.

We believe that one effective method of multivariate time series anomaly detection is mapping the MTS data from hyperspace to low-dimensional subspace with the same assemblage's characteristics, and then, traditional anomaly detection methods can be used in low-dimensional subspace, reducing the repeated computation and improving efficiency by running and filtering method. Another important idea considers the MTS sequence of greatest difference as MTS anomaly, so the MTS anomaly detection method can be obtained by the specific definition of MTS anomaly and the research findings of MTS similarity measurement and index query.

5. Summary and Outlook

In summary, several suggestions are put forward on time series anomaly detection: (1) Anomaly is a relative concept. It is false to consider anomaly as a duality characteristic, either normal or not. (2) The concept and significance of anomaly is different in different areas. For example, it is obviously abnormal when a person's height is 568 cm, but when a company CEO' wages which is far higher than the vast majority of employees, the wages data cannot be considered as abnormal. As a result, the causes of the exception are different, whether the detection algorithm works correctly due to the judgment of experts in this area. The algorithm only to provide users the suspicious data to attract their attention. Therefore, viewed from this perspective, anomaly detection can also be considered as a query with a special condition or meaning. (3) The MTS anomaly detection is largely different from one of traditional multivariate data and UTS. Because MTS has the dual characteristics of multivariable and ordered data. The MTS anomaly detection can be obtained by the comprehensive analysis of each variables sequence usually.

In short, the abnormal detection of time series is not mature, especially in MTS anomaly detection. In addition, the algorithm efficiency of current anomaly detection algorithm is still not satisfactory, so how to further reduce the complexity of the algorithm to fit dynamic time series anomaly mining requires further investigation.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China through grant 61502534, University Foundation of People Armed Police Engineering University of China through grant WJK201511 and the Open Funds for the Joint Laboratory of China satellite measurement and control center through grant FOM2015OF015.

References

[1] Bowen. Z. and Shahriar. S, "Finding needle in a million metrics: anomaly detection in a large-scale computational advertising platform", Proceeding of the 2nd International Workshop on Ad Targeting at Scale, San Francisco, CA, USA, pp.1062-1065, Feb, 2016

[2] Whitehead. B, Hoyt. W. A, "Function approximation approach to anomaly detection in propulsion system test data", Journal of Propulsion and Power, vol.11, no.5, pp. 1074-1076, May, 2015.

[3] Hawkins. D, "Identification of outliers", Chapman and Hall, London, 1980.

[4] Ozkan, Huseyin.O, Fatih. O, Suleyman. S, "Online anomaly detection under markov statistics with controllable Type-I error ", IEEE Transactions on Signal Processing, Vol.64, No.6, pp.1435-1445, March, 2016.

[5] Chandola. V, Banerjee. A, Kumar. V, "Anomaly Detection: A Survey", ACM Computing Surveys, Vol.41, No.3, pp. 1-58, Nov, 2009.

[6] Ruts. I, Rousseeuw.P, "Computing depth contours of bivariate point clouds", Computational statistics and data analysis, Vol.23, No.1, pp.153-168, Jan, 1996

[7] Hardin. J, Rocke. D. M, "Outlier detection in the multiple cluster Setting using the minimum covariance determinant estimator", Computational Statistics and Data Analysis, Vol.44, No.4, pp.625-638, Apr, 2004.

[8] Sarawagi. S, Agrawal. R, Megiddo. N, "Discovery driven exploration of OLAP data cubes", Proceeding of the 6th international conference on extending database technology, Valencia, pp.168-182, Mar, 1998.

[9] Jilin Qu. "Indexing and Querying of time series in data

mining", Tianjin University, Tian Jing, 2006.
[10] Markou. M, Singh. S, "Novelty detection: a review part2: neural network based approaches", Sign Processing, Vol.83, No.12, pp.2499-2521, Dec, 2003.
[11] Knorr. E. M, Ng. R. T, "A unified notion of outliers: properties and computation", Proceeding of the 3rd international conference on knowledge discovery and data mining, Newport Beach, pp.219-222, Aug, 1997.
[12] Ramaswamy. S, Rastogi. R, Shim. K, "Efficient algorithms for mining outliers from large data sets", Proceeding of the 2000 ACM SIGMOD international conference on management of data, New York, pp. 427-438, Jul, 2000.
[13] Dazhuo Zhou, "Clustering, similarity search and outlier detection in multivariate time series", Tianjing University, Tianjing, 2008.
[14] Agyemang. M, Ezeife. C. I, "Large scale Mine. Algorithm for mining local outliers", Proceeding of the 15th information resource management association international conference, New Orleans, pp.5-8, Apr, 2004.
[15] Papadimitriou. S, Kitagawa. H, Gibbons. P. B, "LOCI： fast outlier detection using the local correlation ingral", Technical report, pp.2-9, 2002.
[16] Shengyi Jiang, Qinghua Li, Hui Wang, "An enhanced approach for mining local outlier", Journal of Computer Research and Development, Vol.42, No.2, pp.210-216, 2005.
[17] Keogh. E, Jessica. L, "Finding unusual medical time-series subsequences: Algorithms and Applications", IEEE Transaction on Information Technology in Biomedicine, Vol.10, No.3, pp.429-439, Mar, 2006.
[18] Sadik. S, Gruenwald. L, "An adaptive outlier detection technique for data streams", Proceeding of the SSDBM 2011 international conference, Portland, pp.596-597, Jul, 2011.
[19] Shahabi. C, Tian. X, "Tsa-tree: A wavelet-based approach to improve the efficiency of multi-level surprise and trend queries", Proceeding of the 12th international conference on scientific and statistical database management, Washington, pp.55-68, Jul, 2000.
[20] Chandola. V, "Anomaly detection for symbolic sequences and time series data", The University of Minnesota, Minnesota, 2009.
[21] Meiyu Sun, "Research on key issues of stochastic non — stationary time series data mining based on fractal theory", Donghua University, Shanghai, 2009.
[22] Xiaoqing Weng, Junyi Shen, "Identification of outlier patterns in multivariate time series", Pattern Recognition and Artificial Intelligence, Vol.20, No.3, pp.336-342, Mar, 2007.
[23] Xin Wang, "Two-stage outlier detection in multivariate time series", Application Research of Computers, Vol.28, No.7, pp.2466-2469, Jul, 2011.
[24] Quan Li, Xingshe Zhou, "Multivariate time series anomaly detection method based on KPCA", Computer Measurement & Control, Vol.19, No.4, pp.822-825, Apr, 2011.
[25] Naoya. T, Takehisa. Y, "Anomaly detection from multivariate time series with sparse representation", Proceeding of IEEE International Conference on Systems, Man and Cybernetics, San Diego, pp.2651-2656, Oct, 2014.
[26] Baragona. R, Battaglia. F, "Outlier detection in multivariate time series by independent analysis", Neural Computation, Vol.19, No.7, pp.1962-1984, Jul, 2007.