Outlier Detection in Climatology Time Series with Sliding Window Prediction

Manish Mahajan, Santosh Kumar, Bhasker Pant

Abstract*: It is important to identify outliers for climatology series data. With better quality of data decision capability will improve which in turn will improve the complete operation. An algorithm utilising the sliding window prediction method is being proposed to improve the data decision capability in this paper. The time series are parted in accordance with the size of sliding window. Thereafter a prediction model is rooted with the help of historical data to forecast the new values. There is a pre decided threshold value which will be compared to the difference of predicted and measured value. If the difference is greater than a predefined threshold then the specific point will be treated as an outlier. Results from experiment are showing that the algorithm is identifying the outliers in climatology time series data and also remodeling the correction efficiency.

***Index Terms*: climatology data, forecast model, Outliers, sliding window, time series.**

I. INTRODUCTION

The detection and analysis of outliers in time series data is a challenging problem encountered in data mining. Many methods have been devised over time which deal with the problem of unsupervised outlier detection problem [1].Climatology is defined as weather conditions averaged over a long period of time. Climatology data forms the basis for developments of metrology and atmospheric science. Climatology data not only provides daily whether forecasts but it is also processed and programmed to get important information regarding agriculture, industry, defense, hydrology and many more with the help of long term statistics. The Indian Metrological Department (IMD) issue warnings for several natural disasters with the help of these processed information. The data regarding rainfall and temperature (maximum and minimum) is of utmost importance for any specific geographical location as with these details one can conclude several parameters. These parameters are very important to characterize thermal status of a particular place. So one has to ensure the accuracy of these parameters. While working with the small amount of data it is possible to handle the anomalies and to manually detect and correct the outliers but with larger streams of data one needs to deals with machines to preserve the accuracy and efficiency.

Time series mining is amongst the exigent problems in the aea of data mining. The data obtained from many a sensors employed in a wide variety of fields all generate a time series data, and analyzing the time series data for the purpose of making intelligent decisions or forecasts necessitates detection of outliers in the data[2]. This paper presents a model (algorithm) for identification of outliers in climatology time series with the help of sliding window prediction. The proposed model can forecast the future values based on the historical datasets. To identify the outliers, a pre-defined value called as threshold value is determined. This value will be checked against the difference obtained between the predicted and the historical value. If there is a difference between these two readings greater than the threshold value then we will call this point as outlier and then it will be corrected. Results are reflecting the fact that this model can be used to identify and correct outliers in climatology time series data successfully.

Remaining sections of the paper are organized as mentioned. In section 2 of this work is about related work to time series data. In section 3 we elaborate the model for identification of outliers in time series. Section 4 is about experiment results and foundation. Section 5 presents the conclusion and possible future work.

II. RELATED WORK

An outlier can be defined as an observation, much deviated from all other points in a test sample. This particular trade of outlier may act for suspicion that it is derived from some other method [3][4]. Anomaly detection and outlier mining are the different terms with almost same meaning. The motive behind outlier detection is to find type of abnormality in any data set and find a way to spot that as an outlier. There are five methods involved in time series anomaly detection.

Detection Method based on clustering:

The data set is divided into several clusters. After this if a point doesn't belongs to any of the above created clusters then it will be stated as an outlier. This method is primarily utilised for unsupervised detection (sometimes also used for for semi supervised detection also) [5] [6].

- **Detection Method based on Density:**

  The points in the data set are assigned a weight and on the basis of that weight it is decided that the particular point is outlier or not. It doesn't give binary result. This works on neighborhood relative theory, so the value depends on the distance of objects to its neighborhood [7] [8].

Revised Manuscript Received on August 05, 2019

Manish Mahajan, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, India.

Santosh Kumar, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, India..

Bhasker Pant, Department of Computer Science and Engineering, Graphic Era (Deemed to be University), Dehradun, India..

- **Detection Method based on Classifier:**

In this method a model has to be established for historical time series. A regression model with the help of support vector regression is established for the same. A new time series and the model has been watched for the result then. A novel application of this approach is a Graph based outlier detection technique[9]. The classifier based outlier detection technique has been successfully applied to WSN data and shown promising results [10].

- **Detection Method based on Fixed Size Windows:**

This method is based on the fact that one big time series can be divided in to fix size small time series which we are calling as window. After this partition we can search for the outliers in each window. So if there is an outlier in the original time series there is a possibility that it may show up in any of the window. The window based method has shown some remarkable levels of performance even when compared with factor based approaches. [11]

- **Detection Method based on Distance:**

The distance based outlier detection techniques were introduced first by Knorr and Ng[12]. According to them "*An object p in a data set DS is an DB(q, dist)-outlier if at least fraction q of the object in DS lie at a greater distance the distance from p".* The results obtained by the simple approach were improved upon by Ramaswamy et al [13] by adding a rank based on the distance and using the rank as a outlier score. The concept of hubness-awareness was introduced in [14] to increase the efficiency of distance based outlier analysis and the results were shown to be promising when applied on multi dimensional data sets. Structural characteristics of the problem also play an important role in predicting the efficiency of performance of the algorithm. The metadata of the problem can be extracted and used as an important parameter towards the outlier analysis and detection approach [15].

The series can be seen as a group of feature points in this method. A multi order model with regression is then used to get the unequal division in the given series.

The biased scores are then calculated with the help of time wrapping. Based on these score one can say a point is outlier or not.

Based on the above methods the performance of detection method based on window depends completely on the size of window. A larger size window or a very small window size can affect the efficiency of the method and the accuracy of the result.

The method based on clustering completely depends on the number of identified clusters and the availability of outliers in the available data set. SVM based methods are popular and are widely in use for outlier detection. Apart from that method based on density have quite a high complexity to work with.

We propose a method to identify outliers in climatology time series data. The method will part the climatology time series into sequences in part with the help of sliding window.

In order to predict value for the next data point in series we have used a prediction model [16]. The threshold value [TV] has been calculated from nearest data point in the data set. If TV shows that our predicted value is showing deviation from it then the point will be treated as an outlier. With the help of prediction model which is based on time series model, it is easy to forecast predictions based on series. Since it is true that the nearest the points the higher the correlation between those points, so it is the best choice to calculate TV and to compare it with the predicted values. Apart from that the value of TV can be determined dynamically based on window.

The performance of the underlying Data mning algorithms can be improved by application of Nature inspired MetaHeuristic algorithms [17][18].

III. PREDICTION WITH SLIDING WINDOW FOR OUTLIER DETECTION IN SERIES

In a time series data points are graphed in accordance with time. In Climatology time series, value of two factors (rain and temperature) recorded as the time varies.

Climatology Time Series Definition:

Climatology time series is seen as collection of ordered pairs of points.

$$L = \{l_1 = (v_1, t_1), l_2 = (v_2, t_2), ...\}$$

Here points $l_i = (v_i, t_i)$ denotes the pointed value v_i at time t_i .

For outlier detection in climatology time series first of all we need to know what points in data are abnormal. One can observe that the changes in climatology time series data are prudent. If we take temperature from climatology data then the changes in temperature with time are discreet most of the time. In very few cases there will be a rapid spike of change in temperature (these may be some unexpected cases).

In climatology time series if one sought to find abnormality then we have to find k-nearest neighbor of a particular point in the time series itself. consider we are taking temperature factor into consideration in the climatology time series data.

Let l_i represents a point in temperature time series $l_i = (v_i, t_i)$ represents temperature v_i at some point of time t_i .

To represent k-nearest neighbor window of our particular point l_i we need to find:

$$H_i^{(k)} = \{l_{i-2k}, l_{i-2k+1}, ..., l_{i-1}\}$$

Now after checking the difference between the actual point l_i and the difference between the values given by our k-nearest neighbor model one can identify a point as an outlier.

Framework for Outlier Identification:

To identify the outlier in climatology time series with the help of sliding window one have to find k-nn window for a particular point l_i . Providing as input the view points of $H_i^{(k)}$ to the prediction model to get v_i (we called it as *'a'* in framework) of point l_i . Afterwards we will be calculating confidence area of l_i analogous to the value of v_i .

While talking about our TV value, it depends on the width of our sliding window (k) and confidence value. So it can be calculated accordingly.

Model for Prediction of Outlier:

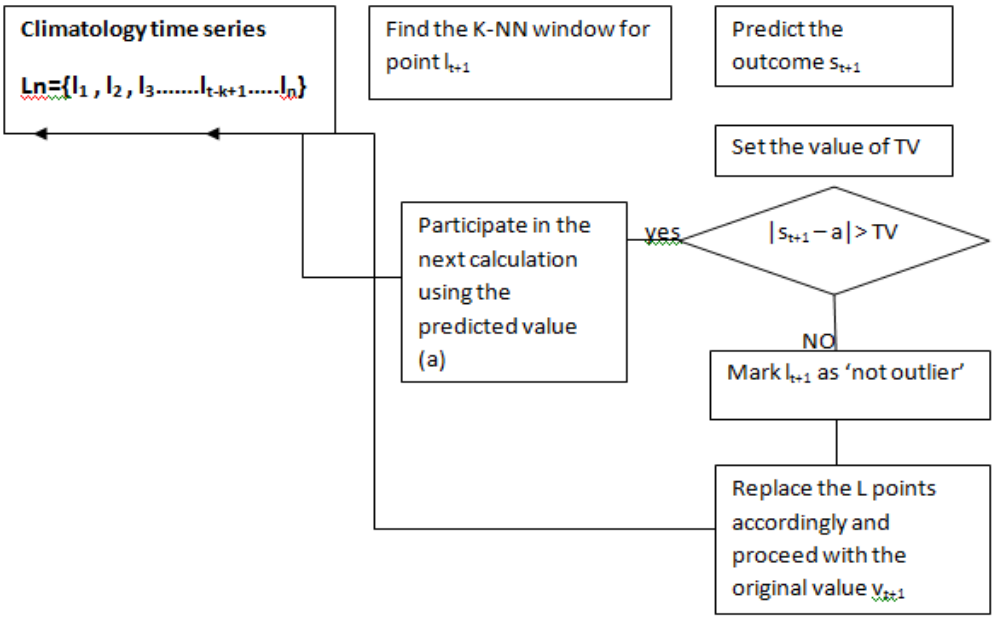
The proposed framework is based on the prediction model. The climatology time series is used as an input to the prediction model. The input parameter for prediction in our frame work is sliding window. When the sliding window $\{l_1, l_2, l_3, \ldots\ldots\ldots, l_i\}$ or we can say $H_i^{(k)}$ is provided to the prediction model then our model can be viewed as:

$$L_{i+1} = A\left(h_i^{(k)}\right)$$

Here $A(\ )$ is our prediction model.

The above model for prediction is statistical. It works on previous performance of variables to predict the upcoming performance.

Fig 1: Framework for identification of outlier in climatology time series with sliding window



Station Name	Month	Period	No. of Years	Mean Temperature in degree C - Maximum	Mean Temperature in degree C - Minimum	Mean Rainfall in mm
Abu	January	1901-2000	100	19.3	8	5.3
Abu	February	1901-2000	100	21	10	4.4
Abu	March	1901-2000	100	25.3	14.5	6.5
Abu	April	1901-2000	100	29.4	18.7	2.6
Abu	May	1901-2000	100	31.5	21	16.4
Abu	June	1901-2000	100	29.1	19.8	101.6
Abu	July	1901-2000	100	24.5	18.7	573.2
Abu	August	1901-2000	100	22.7	17.8	600.3
Abu	September	1901-2000	100	24.5	17.6	214.2
Abu	October	1901-2000	100	26.7	16.2	19.4
Abu	November	1901-2000	100	23.8	12.1	7.9
Abu	December	1901-2000	100	20.9	9	2.4
Agartala (A)	January	1953-2000	43	25.6	10	27.5
Agartala (A)	February	1953-2000	43	28.3	13.2	21.5
Agartala (A)	March	1953-2000	43	32.5	18.7	60.7
Agartala (A)	April	1953-2000	43	33.7	22.2	199.7
Agartala (A)	May	1953-2000	43	32.8	23.5	329.9

Table : Sample Data from the Data Set by IMD

Selection of the Parameters:

To effectively identify the outliers in the climatology time series, the method to find unusual points in the series which works on sliding window concept, needs to identify proper threshold for all the test points. That is why the two parameters p and k in the framework are the most important issue for the betterment of the overall process.

We selected the parameters according to the below mentioned rules:

- K is the width of the window. It plays an important role as a smaller or to large window size can affect the results. If the value of k is larger than the correlation between the participating points will be higher. This in turn will increase the complexity for computation. For getting this thing on optimal range we suggested to vary k in between 2 to 16.

  K= {2, 3, 4.....16} with an increment of 1

- P is the con.coefficient. What is the probability that the values measured are going to be in the specified interval is provided by this coefficient p. The range of confidence interval depends on the value of confidence coefficient (larger the coefficient greater the range). P will grow from 82% to 100 % with an increment of 3.

  P= {82, 85, 88....100}

IV. ANALYSIS BASED ON EXPERIMENTS

Our data set contains climatology data of few Indian cities. It is provided by the meteorological department of India. The data is segregated by city name, year and mean minimum and maximum temperature. It is released under data sharing and accessibility policy (NDSAP). It contains the records from year 1901 to 2000. A sample of the data included in the dataset is given in the table.

Result Evaluation:

The sliding window width is 5 and the coefficient of confidence value is 85%.

The red dotted lines in fig 3 are confidence area. Blue plot is the actual value and yellow plot is predicted value. The black dots on the confidence boundary represent the outliers. Below is the result for the efficiency of the framework to detect the outliers.

Fig 2 : Outlier Detection Algorithm Plot

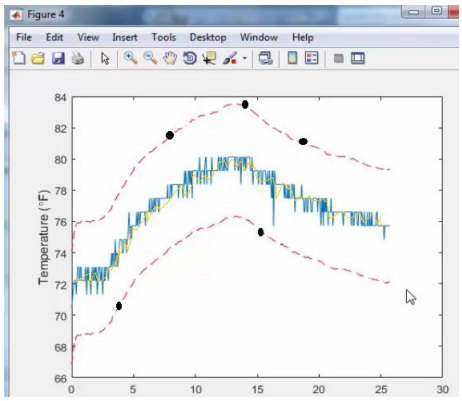
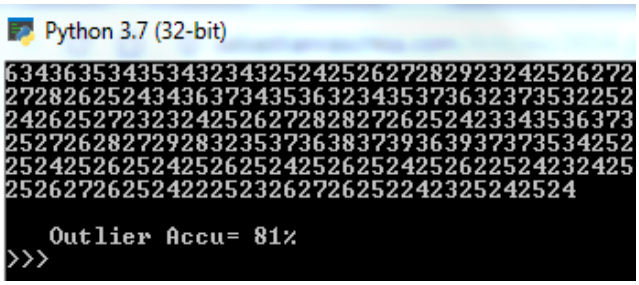


Fig 3: Accuracy of the proposed algorithm



V.CONCLUSION

Climatology time series shows significant amount of data which is used and further can be used in many other models. With the help of sliding window method we have shown how to identify the outliers in the climatology time series.The accuracy for identifying the outliers with this frame work is 81%.

There is much more work left in the area of time series data. There is probability that the external source from where the time series is generating the data might get corrupted. In that case it will be interesting to know which sensor node got faulty and how to identify it to recover it. Moreover the role of Nature inspired metaheuristics in optimization of the data mining and outlier detection algorithms needs to be explored in detail.

REFERENCES

1. Campos, G.O.; Zimek, A.; Sander, J. et al.;, On the evaluation of unsupervised outlier detection., Data Min Knowl Disc (2016) 30: 891
2. Nikolay L.; Saeed A.;Flint I. 2015. Generic and Scalable Framework for Automated Time-series Anomaly Detection. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15). ACM, New York, NY, USA,
3. Hawkins, D.M. Identification of Outliers. Biometrics 1980, 37, 860.
4. Aggarwal C. (2015) Outlier Analysis. In: Data Mining. Springer, Cham
5. Jiang, F.; Liu, G.; Du, J.; Sui, Y. Initialization of K-modes clustering using outlier detection techniques. Inf. Sci. 2016, 332, 167–183.
6. Jobe, J.M.; Pokojovy, M. A Cluster-Based Outlier Detection Scheme for Multivariate Data. J. Am. Stat. Assoc. 2015, 110, 1543–1551.
7. Liu, J.; Deng, H.F. Outlier detection on uncertain data based on local information. Knowl.-Based Syst. 2013, 51, 60–71.

8. Cassisi, C.; Ferro, A.; Giugno, R.; Pigola, G.; Pulvirenti, A. Enhancing density-based clustering: Parameter reduction and outlier detection. Inf. Syst. 2013, 38, 317–330.
9. Akoglu, L.; Tong, H.; Koutra, D., Graph Based anomaly detection and description: a survey, Data Min Knowl Disc (2015) 29: 626.
10. Titouna, C.; Aliouat, M.; Gueroui, M.;, Outlier detection approach using Bayes classifiers Wireless Pers Commun (2015) 85: 1009.
11. Zhang L; Lin J.; Karim,R.; Sliding Window-Based Fault Detection From High-Dimensional Data Streams, in *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 47, no. 2, pp. 289-303, Feb. 2017.
12.]Knorr, E.M. ; Ng, R.T. Algorithms for mining distance –based outliers in large datasets. In Proc. 24th Int. Conf. Very large Data Bases, VLDB Pg. 392-403, 1998
13. Ramaswamy, S; Rastogi, R; Shim, K. Efficient algorithms for mining outliers from large data sets. Proc. Of ACM SIGMOD Int. Conf. on Management of Data, 2000, Pg. 427 – 438
14. Flexer A, An Empirical Analysis of Hubness in Unsupervised Distance-Based Outlier Detection, 2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW), Barcelona, Spain, 2017, pp. 716-723.
15. Ferrari, D.G.; Castro, L.N.; Clustering algorithm selection by meta-learning systems: A new distance-based problem characterization and ranking combination methods, Information Sciences, Vol 31, April 2015, Pg 181-194
16. Zhao, N.; Yu, F.R.; Sun, H.; Yin, H.; Nallanathan, A.; Wang, G. Interference alignment with delayed channel state information and dynamic AR-model channel prediction in wireless networks. Wirel. Netw. 2015, 21, 1227–1242.
17. Mahajan M., Kumar S., Pant B. (2019) A Novel Cluster Based Algorithm for Outlier Detection. In: Iyer B., Nalbalwar S., Pathak N. (eds) Computing, Communication and Signal Processing. Advances in Intelligent Systems and Computing, vol 810. Springer, Singapore
18. Kant N., Mahajan M. (2019) Time-Series Outlier Detection Using Enhanced K-Means in Combination with PSO Algorithm. In: Ray K., Sharan S., Rawat S., Jain S., Srivastava S., Bandyopadhyay A. (eds) Engineering Vibration, Communication and Information Processing. Lecture Notes in Electrical Engineering, vol 478. Springer, Singapore

AUTHORS PROFILE

Manish Mahajan (manishmahajan@geu.ac.in) is pursuing Ph.D. from Graphic Era University, Dehradun. He received his B.E (Electronics Telecommn.) from Marathwada University in 1992 and M.tech in 2005. He is a member of ACM, SCRS, IAENG, and has published 14 research papers in National and International Journals/Conferences in the fields of Computer Security, Software Testing and Data Mining. He has over 20 years of experience in teaching/research of UG(B.Tech) and PG(M.Tech) level courses as a Lecturer/Assistant Professor/Associate Professor in organizations like ABES IT, Ghaziabad, IPEC, Ghaziabad, ABES EC Ghaziabad and Graphic Era University, Dehradun to name a few. He is currently associated with Graphic Era University as Associate Professor and is Heading the of Department of Information Technology. His research interests include Data Mining, Outlier Analysis, Software Testing, Cyber and Information Security.

.

Santosh Kumar (amu.santosh @ gmail.com) received Ph.D. from IIT Roorkee (India) in 2012, M. Tech. (CSE) from Aligarh Muslim University, Aligarh (India) in 2007 and B.E. (IT) from C.C.S.University, Meerut (India) in 2003. He has more than 13 years of experience in teaching/research of UG (B. Tech.) and PG (M.Tech.) level courses as a Lecturer/Assistant Professor/ Associate Professor in various academic /research organizations. He has supervised 01 Ph.D. Thesis and 22 M.Tech Thesis and presently mentoring 06 Ph.D. students (singly and jointly) and 03 M.Tech students. He has also completed a consultancy project titled "MANET Architecture Design for Tactical Radios" of DRDO, Dehradun in between 2009-2011. He is an active reviewer board member in various national/International Journals and Conferences. He has memberships of ACM (Senior Member), IEEE, IAENG, ACEEE, ISOC (USA) and contributed more than 52 research papers in National and International Journals/conferences in the field of Wireless Communication Networks, Mobile Computing and Grid Computing and software Engineering. Currently holding position of Associate professor in the Graphic Era Deemed to be University, Dehradun (India). His research interest includes Wireless Networks, MANET, WSN, IoT, and Software Engineering..

Bhasker Pant (pantbhasker2@gmail.com)
Currently working as Dean Research & Development and Associate Professor in Department of Computer Science and Engineering. He is Ph.D. in Machine Learning and Bioinformatics from MANIT, Bhopal.
Has more than 15 years of experience in Research and Academics. He has till now guided as Supervisor 3 Ph.D. candidates (Awarded).and 5 candidates are in advance state of work. He has also guided 28 MTech. Students for dissertation. He has also supervised 2 foreign students for internship.Dr.Bhasker Pant has more than 70 research publication in National and international Journals.
He has also chaired a session in Robust Classification & Predictive Modelling for classification held at Huangshi, China.

.