ISSN: 2277-9655
(I2OR), Publication Impact Factor: 3.785

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A SURVEY ON OUTLIER DETECTION FOR UNCERTAIN METEOROLOGY DATA

K.Masila Devi*, R.Karthikeyan
Department of Computer Science and Engineering, PSNA College of Engineering and Technology, Dindigul, Tamilnadu-624619, India

ABSTRACT

In recent times climate change has been considered to be the major problem. Meteorology is the study of atmosphere like humidity, pressure, etc. It is important to analyze the hidden knowledge about the huge meteorological data in order to predict the climate from affecting various sectors. So the large data is analyzed, clustered, outliers are removed to predict the climate condition. As the weather data are large and continuous there are chances of many anomalies to be present. In order to remove those anomalies and predict the climatic condition a special outlier analysis is undergone. In this paper, a survey for detecting outlier is discussed.

KEYWORDS: Applications, Clusters, Outlier detection.

INTRODUCTION

Outlier detection is to find patterns that do not match to expected behavior. Outliers are patterns in data that do not belong to a well defined notion of normal behavior. Also an outlier in a set of data that appears to be different or abnormal with the remainder of the data in the set. Outliers arise due to many reasons such as human error, instrument error, natural deviations in populations, fraudulent behaviour, changes in behaviour of systems or faults in systems, etc. An outlier can also be either a single data point or a cluster of data.

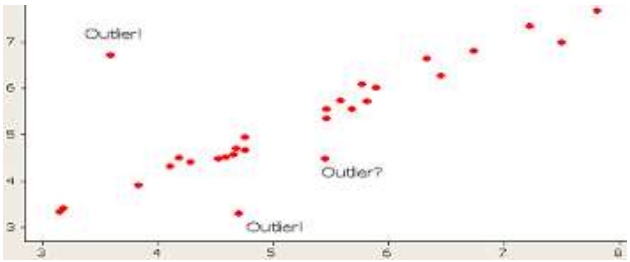

Fig 1: Outlier detection

The approaches used in the problem of outlier detection are:

1. To determine the outliers with having no prior knowledge about the data. This is essentially a learning approach similar to unsupervised clustering.
2. To model both normality and abnormality of the data. This approach which is similar to supervised classification and requires pre-labeled data and tagged as normal or abnormal data.
3. It is similar to a semi-supervised recognition or detection of data and can be considered semi-supervised as the normal class is taught but the algorithm is used to recognize the abnormality.

Outlier detection has numerous applications, including detection of credit card fraud, to find criminal activities in E-commerce, video surveillance, pharmaceutical research analysis, prediction of weather, the analysis of performance statistics of professional athletes and discovery of unexpected or unknown astronomical objects or phenomena. Weather condition is described as the state of the atmosphere at a given time and place. Weather forecasts are done by collecting numerous data about the current state of the atmosphere. Weather forecasting is used to determine how the present state of the atmosphere will change accordingly.

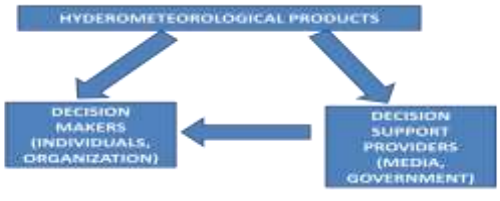

Fig 2: weather prediction

There are several issues that arise in prediction are dimensional characters, redundancy of data, missing data, skews in the data, invalid data etc. In uncertain data management, data records are typically represented by probability distributions rather than deterministic values. Here detecting rare patterns in climatic change using mining techniques is used to understand and detect the patterns of change in climate based on the probabilistic values.

LITERATURE REVIEW

The survey on many outlier detection in data streams and uncertain data are studied here. Initially detecting outliers in large data set is a difficult task. One method that is been carried out to detect outliers in large data set is, A method for detecting and deleting distance based outliers in very large data sets introduces parallel computation so as with spare additional time and having magnificent execution. Initially, a graph is made based on the data set. Then weights are assigned to each of the data's in the graph. Algorithm proceeds by assigning weights to the data in each row. As per the weight a threshold value is set. If there is a chance that data exceeds the threshold value, then that row is considered as an outlier and it is removed from the data set. In this way outliers from all the data sets are obtained. After deletion of the particular row more data's can be added to the data set. Automatically weights are assigned. And the same procedure will be repeated. By deleting the outliers, it increases the space for storing more data [4].

Then for detecting outliers in large database a new algorithm was introduced. The technique reports all outliers by scanning the dataset at most twice. So in this paper a new algorithm, SNIF (scanwith prioritized flushing) was introduced. Afford to hold more objects in memory during the first dataset scan itself, which allows permits a significant portion of R as non outliers straightforwardly after the scan. As a result, the remaining objects that require further verification may fit in memory, so that another scan of R suffices to determine the exact outliers. Based on this idea, SNIF deploys a novel prioritized flushing technique to minimize the shot of performing the third scan of R. Specifically, the technique relates every object with a "priority", and, whenever the memory becomes full, flushes the objects with the lowest priorities [10].

BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies), and shows that it is particularly suitable for very large databases. This system finds a good quality clusters in a single scan of data. Balanced Iterative Reducing and Clustering using Hierarchies is introduced for very large databases. But BRICH cannot increase the threshold dynamically. And not possible to dynamically adjust outlier criteria[8]. For weather prediction the data will be continuous and sometimes uncertain. So in that case these techniques do not support well. So we need to concentrate on the outlier detection for continuous and uncertain data streams since weather data are continuous and uncertain.

Collecting large amount of data from the online transactions datasets which is not possible to monitor and store in the memory. To rectify this problem Sliding window method is implemented to allocate the memory space for the dataset.

Anamoly detection in data stream, Angle based system is used to discover the outliers in diverse techniques which is more effective. Index based algorithm and Random hyper plane projection algorithm are the two distinct algorithms used as the part of Angle based method to find the outliers. Index based algorithm is set to find indexing value between the objects. Random hyper plane method is used to find the exact value in the dataset randomly. Sliding window method is implemented to allocate the memory space for the dataset[2].

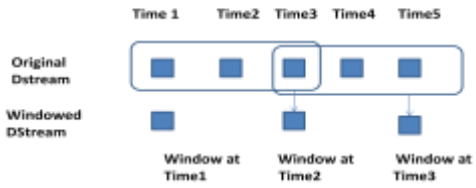

Fig 3: Sliding window for data stream

An iterative random sampling procedure is developed that examines apart of the entire data space. If an object does not belong to the examined space , it is considered an outlier in the relating iteration. In other words ,if an object is in that space ,it is viewed an inlier .By repeating this procedure of space examination ,the proposed strategy figures the inlierness score, which is called the OF(Observabilityfactor) ,for each object in a dataset. Since the OF indicates the inlierness score, an object with a low OF value is a promising outlier candidate hopefully. This method is related to distance-based approaches, because it indirectly utilizes the distances between objects[9].
Numerous outlier detection techniques require user-defined parameters which requires area knowledge for the effective operation over data streams. Because of the dynamic nature of data streams, the parameter values are very hard to predict and the parameter that

is chosen may not be suitable throughout the lifetime of a data stream

Since the weather data will be uncertain it is important to detect outliers in such data stream. Outlier detection on uncertain data streams which can enable the concurrent execution of different queries simultaneously. Even though it reduces the required storage overhead, More efficient, Offer significant flexibility with regard to the input parameters[1].

Persistent uncertain outlier detection which can rapidly decide the way of the indeterminate components by pruning to enhance the effectiveness. Pruning is to diminish the identification cost. It is an expected outlier probability. Outlier detection on continuous data streams which can deal with the parameter variable queries simultaneously. But Only the changes of k in (R, k,lambda) is studied this paper[3]. Numerous procedures depend on the essential attributes of a particular data to distinguish the outilers. So they can't be effectively connected to diverse sorts of data in other application spaces. They ought to be tuned and modified to adjust to the new space. In this paper a graph based methodology for the disclosure of contextual outlier in sequential data is proposed. The algorithm offers a higher level of adaptability and requires less measure of data about the nature of the analyzed data compared to previous methodologies. The accuracy and efficiency highly relies on the input parameters and thresholds of the data stream. Therefore Cannot automatically estimate the values for different datasets [5].Initially, a cell-based approach of distance-based outlier detection on uncertain objects following by the Gaussian distribution is proposed. Second, an approximate cell-based approach of outlier detection using the bounded Gaussian distribution is proposed to increase the efficiency of outlier detection. Approximation in the Gaussian distribution outperforms the bounded Gaussian distribution enabling more effective pruning. And therefore the degree of uncertainty is bounded [7]. Mining Outlier in Data Streams Using Massive Online Analysis Framework. requires only Very less time to execute. But only the traditional methods are verified using this framework[6].

UNCERTAIN WEATHER DATA

Lately numerous procedures are utilized for collecting data which prompts the uncertainty in data. The primary areas of exploration are,

- Uncertain data Modeling.

A key issue is modeling uncertain data. The complexities are recognized and uprooted that makes the data to be helpful for database management applications.

- Managing Uncertain data.

For this situation, the traditional database management techniques for uncertain data are being carried out. Such as join processing, query processing, indexing, or database integration and other database techniques.

- Uncertain data mining.

The data mining applications are influenced because of the vulnerability in the uncertainty in the data. So, it is not possible to design data mining techniques with such uncertainty during the computations.

The Meteorology data consists of much uncertainty. So detecting outliers in such data is a difficult task because of the streaming nature of the data.

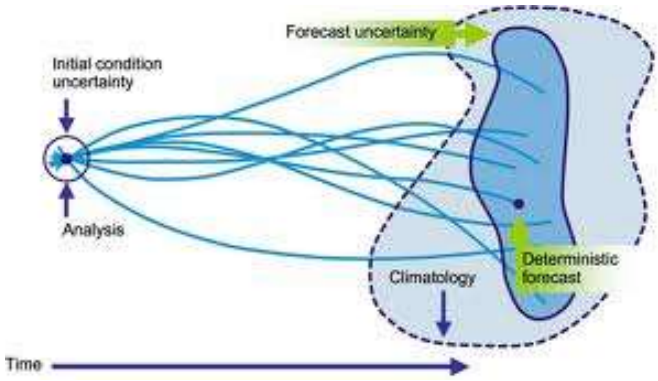

Fig 4: uncertain data

Numerous applications deals with the two sorts of uncertainty. They are:

1) Existential uncertainty: Here, a data may or may not present in the dataset but the presence or absence of such data may influence the likelihood of the other data's presence or absence.

2) Uncertainty over attribute: Here, a number of tuples and the model of that tuples are already determined. The uncertainties of the individual attributes are modeled using the likelihood function, or any parameters like variance.

In the weather data the presence of any data may influence the presence or absence of other data. So the prediction of weather in such condition is a difficult task.

PREPROCESSING AND CLUSTERING

An important step in the data mining process is data preprocessing. It is the initial step too. The challenges in meteorological data quality of the data. This is the reason that makes the data to get exact and redress

forecast of climatic condition. First we have to choose only the related attributes for mining task. So we neglect many attributes like year, wind bearing and time of the most astounding wind, speed, and so on. Then filling of the missing data with the most appropriate values like either zero substitution or mean value substitution. As we are working with weather data that is a form of time series, smoothness and consistency of the series in the dataset have to be maintained.

Clustering has been applied in much meteorological application like determinate the precipitation weather type by finding the similarity between satellite cloud images, seasonal clustering and climatology. Many clustering algorithms like k-means, KNN can be used if we are about to cluster the data using distance based algorithm or probability based algorithms can also be used. In case we use k-means clustering algorithm as it is the most popular clustering algorithm used in scientific and many industrial applications each of the k clusters are formed using the mean values that is called as centroid. In this way all the data are clustered. Each cluster represent the climatic condition for example, minimal measure of downpour, higher temperature, higher dampness rate and slower wind rate etc. The understanding of seasons is very much important to many sectors like agriculture, vegetation, water resources and tourism which largely dependent on the weather conditions.

SLIDING WINDOW

Typically data stream is a period arrangement and may be infinite. So to perform outlier detection in such a stream is an essential fact. To do so, a sliding window technique is utilized. In this, we are supposed to consider n most recent data elements. And the data elements within the window are considered to be active and data mining techniques are applied to it. As the time progresses, new data element arrives, and old data element expires. Here the deletion of old data element is implicit. The window consider the current data elements based on the time or the number of data elements to be considered per slide (ie) A window, W, can be either time-based or count-based. The window is triggered continuously for the rest of the data stream and work in the similar fashion. The important factor to be consider is, the window size. Because it plays a vital role in data stream manipulations. As the meteorological data are continuous we are supposed to use sliding window to process the n recent data. Choosing an eminent window size will lead to effective processing. The size of the sliding window can be decided according to the applications and the system resources. The recently generated transactions in the window will influence the mining result of the sliding windowing,

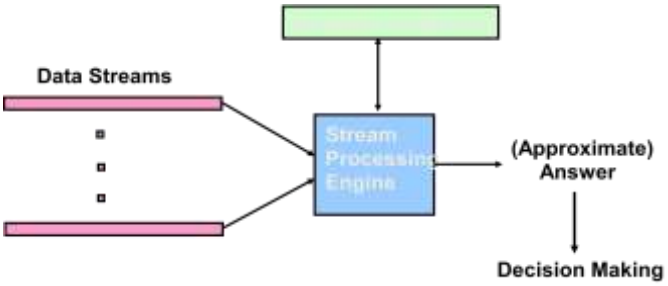

Fig 5: sliding window

The count-based window can also be captured by a time-based window by assuming that a uniform number of transactions arrive within each time unit. In meteorological data may use both type of window as time based.

OUTLIER DETECTION

Outliers in weather data can occur due to the error in entering the data and faulty data collection by the instruments, or it can also represent the abnormal change or any sudden natural events such as tornadoes, hurricane, and forest fires, etc. The diagram below shows the anomaly in temperature data.

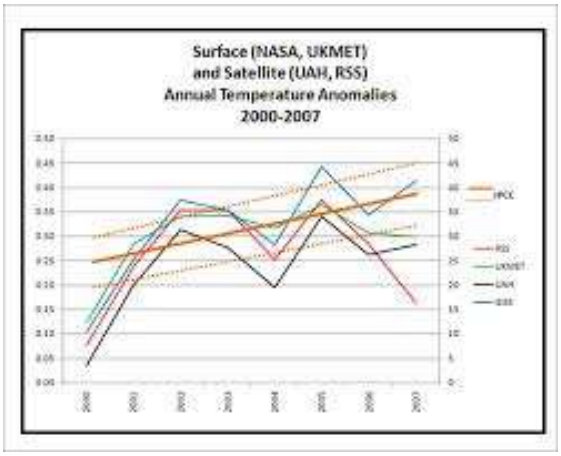

Fig 6: Anomaly in weather data

The outliers are not just a single data that deviate from the data set. But it can also be a cluster of data that totally deviate from the data set.

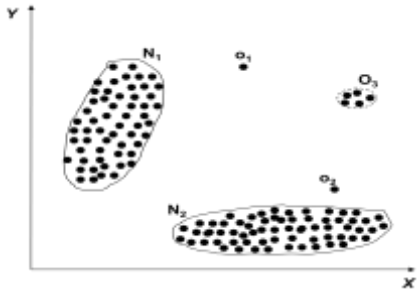

Fig 7: Outlier

Computation cost for evaluation of outlier in dynamic data is high. Data is passing with a time constraints in distributed environment only one pass of scan is possible so we required much memory for data analysis and storage. So it require that minimize the time consumption while using a reasonable amount of memory for storage and computational tasks. One of the major problem in the existing approach is that handling dynamic change in data where it is hard to identify data stream distribution in prior. Even direct computation of probabilities is difficult. In the meteorological data if there is an existentialprobability the outlier detection is different. Because the presence or absence of a data will influence other data.

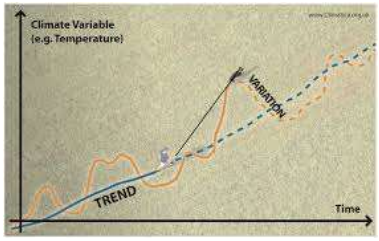

Fig 8: Outlier in weather data

So the variation is alone not an outlier in the case of existential probability. The probability of each data is calculated and then the anomaly is detected.

WEATHER PREDICTION

Now the Classification is utilized in numerous meteorological applications; for example to predict the weather on a particular day will be "sunny", "rainy" or "cloudy" based on the classification. Also it used widely to classify geographical location based on its climate or atmospheric characteristics and classify weather conditions based on the agricultural crops suitable to cultivate on each climate. So in such a case the probability has a huge impact. As there are chances of climatic change due to the other data. For example, consider a cluster have been formed based on the temperature like hot, warm, cold. Now if any is supposed to be absent in a cluster then that cluster would lead to different prediction of weather while classifying. We can use any of the classification algorithms.

OTHER APPLICATIONS

There are numerous applications for detecting outliers in uncertain data streams.

Intrusion Detection Systems: **N**etwork traffic and movement, or other activity in the system may demonstrate some unusual behavior . The detection of such movement is referred to as intrusion detection.

Credit Card Fraud: Credit card fraud is quite predominant. This normally prompts to unauthorized use of the credit card and it is recognized as outliers in credit card transaction data.

Medical Diagnosis: In numerous therapeutic applications the data is gathered from a variety of devices such as MRI scans, PET scans or ECG time-series. Unusual patterns in such data typically reflect disease conditions.

Earth Science: A significant amount of spatiotemporal data about weather patterns, atmosphere changes, or land cover patterns is collected through a variety of mechanisms such as satellites or remote sensing. Abnormal data in such collection provide significant insights about hidden human or environmental trends, which may have caused such anomalies.

CONCLUSION

The outlier detection in the uncertain data stream having data with underlying probabilities is an extremely crucial component in meteorological data. As many sectors like tourism, agriculture, fishery all relies upon the weather condition, the outliers in such data will lead to wrong prediction of weather. In this paper a detailed survey of various advantages and disadvantages of detecting outliers in large dataset, streaming data, and uncertain data are discussed along with many data mining techniques like outlier analysis, prediction and clustering. This survey helps to provide an understanding of the foundational issues and would serve as a beginning point to experts and analysts in concentrating on the vital and rising issues in this field.

REFERENCES

[1] Maria Kontaki, Anastasios Gounaris , Apostolos N.Papadopoulos, Kostas Tsichlas,Yannis Manolopoulos, "Efficient and flexible algorithms for monitoring distance-based outliers over data streams", Published by Elsevier Ltd.,2015.

[2] Ke-Yan Cao, Guo-Ren Wang, Dong-Hong Han, Guo-Hui Ding, Ai-Xia Wang and Ling-Xu Shi, "Continuous Outlier Monitoring on Uncertain Data Streams",JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 29(3): 436{448 May 2014.

[3] K. Vijay Babu, Dr. R. Shriram ," Outlier detection in data streams", International Journal of Future Innovative Science and Technology (IJFIST), *Vol(xx) Issue-xx, May 2015.*

[4] Nithya.Jayaprakash, Ms. Caroline Mary, "Detection and Deletion of Outliers from Large Datasets". International Journal of Innovative Research in Science, Engineering and Technology *Volume 3, Special Issue 5, July 2014.*

[5] Ali Rahmani , Salim Afra, Omar Zarour, Omar Addam, Negar Koochakzadeh, Keivan Kianmehr, Reda Alhajj, Jon Rokne, " Graph-based approach for outlier detection in sequential data and its application on stock market and weather data", Elsevier 2014.

[6] Prof. Dr. P K Srimani, Malini M Patil, "Outlier Mining in Data Streams Using Massive Online Analysis Framework", International Journal of Conceptions on Computing and Information Technology. *Vol. 3, Issue. 1, April' 2015.*

[7] Salman A. Shaikh, Hiroyuki Kitagawa, "Efficient distance-based outlier detection on uncertain datasets of Gaussian distribution", Springer, 17 April 2013.

[8] Tian Zhang, Raghu Ramakrishnan, Miron Livny, "BIRCH: An Efficient Data Clustering Method for Very Large Databases", SIGMOD ACM 1996.

[9] Jihyun Ha, Seulgi Seok, Jong-Seok Lee, "Precise ranking method for outlier detection", Elsevier 88-107, 2015.

[10] Yufei Tao, Xiaokui Xiao, Shuigeng Zhou, "Mining Distance based Outliers from Large Databases in Any Metric Space.",ACM, August 20–23, 2006.

[11] Miss. Kavita Thawkar,Prof. Snehal Golait, Prof, Rushi Longadge, " A Framework for an Outlier Pattern Detection in Weather Forecasting", International Journal of Computer Science and Mobile Computing, Vol.3 Issue.5, May- 2014, pg. 348-358.