Iterative Local Outlier Detection with Bootstrap Aggregation for Dynamic Uncertain Data

Ramesh Kumar B
Assistant Professor, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India

Aljinu Khadar K V
M.Phil Scholar, Department of Computer Science, Sree Narayana Guru College, Coimbatore, Tamil Nadu, India

Abstract – Detecting outliers is one of the major tasks in data mining techniques. The dynamic and uncertain data in wireless sensor network have a huge concentration of outliers at the time of clustering or classification. This work proposes an effective and memory efficient outlier detection technique that gathers data from sensor networks which include temperature, humidity, pressure, voltage etc. The system investigates and utilizes the characteristic of the uncertain objects to explore the outlier and cluster them accordingly. The proposed system studies about the accuracy related issues in the outlier detection and learning phase of dynamic objects. It also aims to efficiently mine the outliers of uncertain dynamic objects using two different algorithms. Sequential pattern mining has been used to extract sequences of frequent events. To enable the continuous monitoring the data uncertainty, the system introduces a special technique called Bootstrap aggregating algorithm and Adaptive weight based LOF algorithm. Existing solutions on uncertain objects offers only limit accuracy. To address this issue, the proposed system uses effective data mining methods to handle the uncertain objects and its streams. The system proposes a new bound score detection technique with the use of local clustering process. The system also performs the outlier detection test for accurate summarized results.

Index Terms – Data Mining, Outlier Detection, One Class Clustering, Anomaly, Accuracy.

1. INTRODUCTION

Outlier detection has been a widely researched problem and finds immense use in a wide variety of application domains such as credit card, insurance, tax fraud detection, intrusion detection for cyber security, fault detection in safety critical systems, military surveillance for enemy activities and many other areas [1]. Outlier detection aims to find patterns in data that do not conform to expected behavior. It has extensive use in a wide variety of applications such as military surveillance for enemy activities, Outlier detection in cyber security, fraud detection for credit cards, insurance or health care and fault detection in safety critical systems. Their importance in data is due to the fact that they can translate into actionable information in a wide variety of applications. The key challenge of clustering in the uncertain data domain is the huge volume of data. Outlier detection schemes need to be computationally efficient to handle these large sized inputs. An Outlier can be an observation that is distinctly different or is at a position of abnormal distance from other values in the uncertain dataset. In many data study tasks a large number of variables are being recorded or sampled for one class clustering. One of the first steps towards obtaining a logical analysis is the detection and expanding observations. Although Outlier is often considered as a new event, they may carry important information. Detected those important or unusual information is more important. It is more important to identify them prior to modeling and analysis. The main aim of the proposed system is discovering exceptional attributes and values by expanding population using popular and effective detection strategies for effective Outlier detection is the main objective of the proposed system. Another problem with Outlier detection techniques is that an increase in uncertainty results in difficulty in identification of Outlier in the full dimensional case [2]. This work aimed at providing a contribution toward the design of automatic methods for the discovery of properties characterizing a small group of Outlier individuals as opposed to the whole population of "normal" individuals.

2. PROBLEM DEFINITION

A key problem in data mining is that of one class learning and active training process for detecting outliers [3]. The problem of one class learning has been widely studied in the data mining community. The addition of noise to the data makes the problem far more difficult from the perspective of uncertainty. It is easy to obtain one class of normal data, whereas collecting and labeling abnormal instances and subset of the collected label may be expensive or impossible. There are several clustering techniques [4] have been applied Clustering algorithms, which are optimized to find clusters rather than outliers. The existing systems [5] suffer from the following basic drawbacks.

- Accuracy of one class learning and outlier detection depends on how good the clustering algorithm captures the structure of clusters.

- in existing system a set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise/outliers. The existing system discovers attributes or properties based on the given populations which are called as inliers. That needs to solve unsupervised problem which is yet unbalanced data learning problem. And several drawback of the existing approach is, the system considers frequent items are considered as normal behavior. The overall drawbacks of the existing system are described below.

- The existing algorithms suffer from the problem with uncertain objects and data streams.
- The algorithms have created many false reports while finding outliers
- Many existing algorithms have failed to obtain both local and global data behavior.

3. PROPOSED SYSTEM

The system investigated and utilized the characteristic of the uncertain objects to explore the group relationship and clustering according to them. The goal is to efficiently mine the one class clustering and subset classification of uncertain dynamic objects using minor clustering and sequential pattern mining. Sequential pattern mining was used to extract sequences of frequent events. To enable the continuous monitoring the group object uncertainty, the system introduces a special technique called minor clustering and Adaptive weight based LOF algorithm. Several solutions on uncertain objects were implemented, but those methods were accuracy constrained. In order to reduce the false alarm the proposed system used effective data mining methods to effectively handle the uncertain objects and its streams. The proposal introduces a novel framework to obtain uncertain objects and one class clustering with semi supervised learning techniques. In this proposal this address the problem of single-class Learning on uncertain data streams and concept summarization learning of the user from history data streams. This proposes a change detection test which is a technique of Pattern Flow to summarize the object cluster and their interest changes. This also presents a novel framework, called Adaptive weight based LOF algorithm

Contributions:

The proposed framework consists of the following contributions.

1. Bootstrap aggregating algorithm: Bootstrap aggregating is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid over-fitting. Although it is usually applied to decision tree methods, it can be used with any type of method.
2. Adaptive weight based LOF algorithm:
3. Bound score calculation process.

Finally this applies Cluster Ensemble process for the final clustering phase. For improving the accuracy and detecting false alarms the proposed system applies adaptive weight based LOF which is a method of LOF -to test whether there will be any change or not in the transition region.

4. PROPOSED METHODOLOGY

The proposed system implements bootstrap aggregation approach and a dynamic AWLOF algorithm and extend the perspective of that approach in order to be able to deal with groups, or sub types, of anomalous individuals. The proposed system considers a rare event and assumes cluster labels of normal and abnormal individuals are given; here, it would be very useful to single out properties characterizing the abnormal individuals. An exceptional property is an attribute characterizing the abnormality of the given anomalous group (the outliers) with respect to the normal data population (the inliers). If the inliers data’s are not much sufficient, then the system will analyze and cross check the available dataset for further process.

4.1 Bootstrap Aggregation method algorithm

The proposed algorithm adopts a strategy consisting in selecting the relevant subsets of the overall set of conditions.

```
Step 1: Read dataset from uploaded data
    a) Read the attributes and values from the transaction $T_{N.}$
    b) Every attribute is set into a variable ‘a’
    c) Set of condition is called ‘C’
Step 2: preprocess
Step 3: attribute extraction
    a. Set $C_a$ as conditions -Identify base conditions for every attribute or properties
Step 4: calculate statistics value
        a. Single clustered data set $S_{c.}$
        b. If the property is already in the cluster- find the value
        c. Else if new  attribute perform the following
        d. Find next value
        e. Find in next cluster
Step 5: identify threshold value for each clustered data
Step 6: detect abnormal and normal from the dataset and return results
```

Step 7: perform phase 2

Step 8: perform cluster ensemble process

Step 9: support vector based clustering process

Step 10: read all the stored rule and threshold for each attribute

Step 11: detect normal details and return

The system performs the above steps in the minor clustering process. Finally that will be combined together with the help of cluster ensemble process.

Uncertain Single-class Learning:

In all, the uncertain single-class Learning for uncertain data streams consists of three steps, as illustrated in the part one of Fig. 3.

Step 1: Initially the one class learning framework generates a bound score for each instance in data set based on its local data behavior.

Step 2: this contains the generated bound score into the learning phase to interactively build an uncertain single-class classifier.

Step 3: finally this integrates the uncertain single-class classifiers derived from the current and historical chunks to predict the data in the target chunk.

In the following, it shows the three steps in detail. For simplicity, in detail the bound score determination and uncertain single-class classifier learning for the current chunk Dc. it can generalize them on other chunks in the same way.

The Extended Bootstrap aggregation based LOF Clustering Algorithm:

In this step, this applies extend the dynamic LOF clustering method to cluster the history chunks represented by support vectors into clusters and each cluster denotes one concept of the user. After collect support vectors of uncertain single-class classifier,

Dynamic AWLOF is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The Dynamic AWLOF Algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields. The Dynamic AWLOF Algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

Algorithm:

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

One of the main disadvantages to Dynamic Extended LOF is the fact that you must specify the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance. For example, if you had a group of people that were easily clustered based upon gender, calling the Dynamic AWLOF Algorithm with k=3 would force the people into three clusters, when k=2 would provide a more natural fit. Similarly, if a group of individuals were easily clustered based upon home state and you called the Dynamic AWLOF Algorithm with k=20, the results might be too generalized to be effective.

Final classifier:

Since one-class data stream learning always extends the one-class methods developed for the static data, the previous work can be classified into two categories. For the methods in the first category they are developed for document-related one-class problems. They extract negative samples and its sub types from the unlabeled examples and then construct a binary classifier by target samples and the extracted negative samples.

The above diagram represents the proposed system. Initially the local data behavior has been collected then the behavior identifies the positive and negative scores along with the sub type. Moreover, each property can have associated with the minor clustering phase with a condition specified and this helps to find the exceptional event from the dataset. In the first step, represent each data chunk using the support vectors of the uncertain single-class classifier built on the corresponding data chunk. In the second step, cluster the history chunks represented by support vectors into clusters and each cluster denotes one concept and summarize the concept of the user.

5. EXPIREMENTS AND RESULTS

This chapter describes the experiments, datasets involved in the experiments. The performance of the proposed system is compared with the existing data one class classification tools.

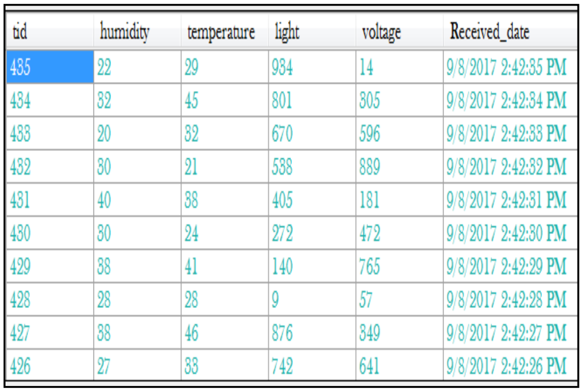
It presents the experimental results for Adaptive weight based LOF algorithm over several datasets which are described below. Some links may provide complete Outlier detection and label availability and some links are limited to the extraction process. Proposed system is implemented in C#.net.

Sensor:

This data stream contains information (temperature, humidity, light, and sensor voltage) collected from 54 sensors deployed in Intel Berkeley Research Lab. The whole stream contains consecutive information recorded over a 2 months period. The same as previous work, this uses sensor from four regions which are covered by ellipses, respectively6. The stream has four classes, in which each region data denotes one class, with 1,051,229 samples, three features.

Synthetic Dataset:

The system created some dynamic synthetic dataset from the above sensor dataset model. The dataset is as follows.



tid	humidity	temperature	light	voltage	Received_date
435	22	29	934	14	9/8/2017 2:42:35 PM
434	32	45	801	305	9/8/2017 2:42:34 PM
433	20	32	670	596	9/8/2017 2:42:33 PM
432	30	21	538	889	9/8/2017 2:42:32 PM
431	40	38	405	181	9/8/2017 2:42:31 PM
430	30	24	272	472	9/8/2017 2:42:30 PM
429	38	41	140	765	9/8/2017 2:42:29 PM
428	28	28	9	57	9/8/2017 2:42:28 PM
427	38	46	876	349	9/8/2017 2:42:27 PM
426	27	33	742	641	9/8/2017 2:42:26 PM

Fig 1.0 Synthetic dataset

The comparative study states that the overall performance of Dynamic AWLOF Algorithm shows outstanding results compared to LOF Algorithm.

Computation Time Analysis and Comparison:

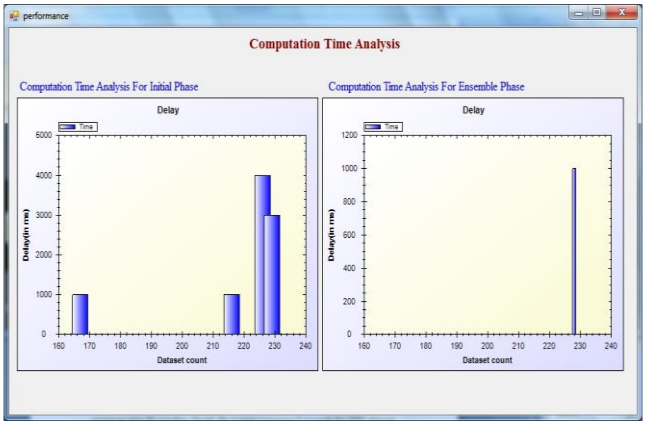


Fig: 2.0 Computation time analyses

Computation memory comparison:

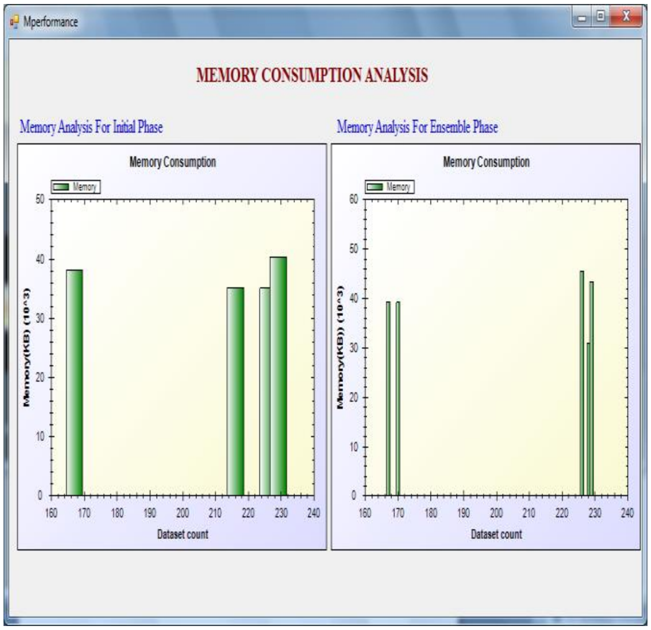


Fig 3.0 memory consumption analysis

Fig: 3.0 show the efficiency comparison between existing two clustering phases in the above diagram. The efficiency in the form of memory is effective than the existing system.

6. CONCLUSION

Finding subset with optimal labeling in the clustering method needs more training data and computational process. The proposed system introduced a new framework for outlier detection subset learning with accurate labeled process. Concept summarization and one class learning with adaptive memory effective LOF will improve the accuracy of the detection method. The proposed framework consists of four parts. The first part generates bound score to capture the local behavior and uncertainty and then build an uncertain single-class classifier by incorporating the uncertainty information into a one-class subset bootstrap aggregation-based framework. Second, this develops support vectors-based AWLOF method to summarize the concept of the user over the history chunks. Third part is the implementation of vague ensemble method to perform the dynamic clustering process. Finally the system uses change detection test to enhance the prediction accuracy. The proposed system has discussed extended k means clustering technique with vague cluster ensemble algorithm. In future the proposed extended K means can be implemented along with the other un-supervised clustering technique. Hence to overcome these problems of K-means algorithm some indexing technique may be added to k-means algorithm in future. Some other directions of the future work is implementing a dynamic ensemble for high dynamic dataset, which may eliminate the re clustering process.

REFERENCES

[1]. Ben-Gal, Irad. "Outlier detection." *Data mining and knowledge discovery handbook* (2005): 131-146.

[2]. Angiulli, Fabrizio, and Clara Pizzuti. "Fast outlier detection in high dimensional spaces." *European Conference on Principles of Data Mining and Knowledge Discovery*. Springer, Berlin, Heidelberg, 2002.

[3]. Hodge, Victoria, and Jim Austin. "A survey of outlier detection methodologies." *Artificial intelligence review* 22.2 (2004): 85-126.

[4]. Amini, Amineh, Teh Ying Wah, and Hadi Saboohi. "On density-based data streams clustering algorithms: a survey." *Journal of Computer Science and Technology* 29.1 (2014): 116-141.

[5]. Ramesh Kumar B. 1. , Aljinu Khadar K V. " A Survey on Outlier Detection Techniques in Dynamic Data Stream" International Journal of Latest Engineering and Management Research (IJLEMR) 2017 23: 30.