INCREMENTAL METHODS FOR DETECTING OUTLIERS FROM MULTIVARIATE DATA STREAM

Simon Fong\textsuperscript{a}, Zhicong Luo\textsuperscript{b}, Bee Wah Yap\textsuperscript{b}, Suash Deb\textsuperscript{c}

Department of Computer and Information Science, University of Macau, Macau SAR, China\textsuperscript{a}
Faculty Computer and Mathematical Sciences, Universiti Teknologi MARA, Selangor, Malaysia\textsuperscript{b}
Department of Computer Science and Engineering, Cambridge Institute of Technology, Ranchi, India\textsuperscript{c}

ecfong@umac.mo

ABSTRACT
Outlier detection is one of the most important data mining techniques. It has broad applications like fraud detection, credit approval, computer network intrusion detection, anti-money laundering, etc. The basis of outlier detection is to identify data points which are “different” or “far away” from the rest of the data points in the given dataset. Traditional outlier detection method is based on statistical analysis. However, this traditional method has an inherent drawback—it requires the availability of the entire dataset. In practice, especially in the real time data feed application, it is not so realistic to wait for all the data because fresh data are streaming in very quickly. Outlier detection is hence done in batches. However two drawbacks may arise: relatively long processing time because of the massive size, and the result may be outdated soon between successive updates. In this paper, we propose several novel incremental methods to process the real time data effectively for outlier detection. For the experiment, we test three types of mechanisms for analyzing the dataset, namely Global Analysis, Cumulative Analysis and Lightweight Analysis with Sliding Window. The experiment dataset is “household power consumption” which is a popular benchmarking data for Massive Online Analysis.

KEY WORDS
Outlier detection; Incremental processing; Data stream mining.

1. Introduction: Background of Outlier Detection Techniques
Numerous researchers have attempted to apply different techniques in detecting outlier, which are generally referred to as defined in the following. “An outlier is an observation that deviates so much from other observations as to arouse suspicions that is was generated by a different mechanism” (Hawkins, 1980). “An outlier is an observation (or subset of observations) which appear to be inconsistent with the remainder of the dataset” (Barnet & Lewis, 1994).

Researchers generally focus on the observation of data irregularities, how each data instance relates to the others (the majority), and how such data instances relate to classification performance. Most of these techniques can be grouped into the following three categories: distribution-based, distance-based, and density-based methods.

1.1 Distribution-based Outlier Detection Methods
These methods are commonly based on statistical analysis. Detection techniques proposed in the literature range from finding extreme values beyond a certain number of standard deviations to complex normality tests. However, most distribution models typically apply directly to the future space and are they univariate. Therefore, they are unsuitable even for moderately high-dimensional data sets. Grubbs proposed a notion which calculates a Z value as the difference between the mean value for the attribute and the query value divided by the standard deviation for the attribute, where the mean and the standard deviation are calculated from all attribute values including the query value. The Z value for the query is compared with a 1% or 5% significance level. The technique requires no pre-defined parameters as all parameters are directly derived from the data. However, the success of this approach heavily depends on the number of exemplars in the data set. The higher the number of records, the more statistically representative the sample is likely to be [1].

In [2], the authors adopted a special outlier detection approach in which the behavior projected by the dataset is examined. If a point is sparse in a lower low-dimensional projection, the data it represents are deemed abnormal and are removed. Brute force, or at best, some form of heuristics, is used to determine the projections. A similar method outlined by [3] builds a height-balanced tree containing clustering features on non-leaf nodes and leaf nodes. Leaf nodes with a low density are then considered outliers and are filtered out.

1.2 Distance-based or Similarity-based Outlier Detection Methods
Distance-based outlier detection techniques are initially introduced by Knorr and Ng [4]. An object \( p \) in a data set \( DS \) is a \( DB(q,\text{dist}) \)-outlier if at least fraction \( q \) of the objects in \( DS \) lie at a greater distance than \( \text{dist} \) from \( p \). This definition is well accepted, since it generalizes
several statistical outlier test. An extension of the above definition was proposed by Ramaswamy et.al. [5]. There is a rank among all the points according to the outlier score. Given two integers kn and w, an object p is said to be an outlier, if less than w objects have higher value for \( D^k \) than p, where \( D^k \) denotes the distance of the \( k^{th} \) nearest neighbor of the object p.

In [6], the researchers first divided data into many subsets before searching for the subset that would cause the greatest reduction in dissimilarity within the training dataset if removed. The dissimilarity function can be any function returning a low value between similar elements and a high value between dissimilar elements, such as variance.

Another team of researchers [7] applied a k-NN algorithm, which essentially compares test data with neighboring data to determine whether they are outliers by reference to their neighbors. In 1936 P.C. Mahalanobis introduced a distance measure [8] which is based on correlations between variables by which different patterns can be identified and analyzed and provides a useful way of determining similarity of an unknown sample set to a known one. It differs from Euclidean distance in that it takes into account the correlations of the data set and is scale-invariant, i.e. not dependent on the scale of measurements. In our paper, we will deploy this method to find outliers.

1.3 Density-based Outlier Detection Methods

In [9], Breunig et al introduced a method that assigns to each object a degree to be an outlier. This degree is called the local outlier factor (LOF) of an object. It is “local” in the manner that the degree hinges on how isolated the object is as to the surrounding neighborhood. In LOF algorithm, observations with high LOF values are regarded as outliers while observations with low LOF values are likely to be normal with respect to their neighborhood. Low-density neighborhood is indicated by high LOF and consequently high potential of being outlier [10]. We will adopt this method in this paper to detect outliers.

2. Methodology of Outlier Detection using Mahalanobis Distance (MD)

2.1 Measurement about Distance

For the earliest statistical-based outlier detection, this method is merely applicable for single dimensional datasets, namely, univariate outliers. In such case, "outliers" in a data set could be done by calculating the deviation for each number, expressed as either a “Z-score” or "modified Z-score" and testing it against certain predefined threshold. Z-score typically refers to number of standard deviation relative to the statistical average. Modified Z-score applies the median computation technique to measure the deviation and in many cases provides more robust statistical detection of outliers.

However, in practice, we usually encounter more complex situations with multidimensional records. One procedure for identifying bivariate outliers and identifying multivariate outliers is called Mahalanobis Distance, and it calculates the distance of particular scores from the center cluster of remaining cases. Formally, the Mahalanobis distance of a multivariate vector \( x = (x_1, x_2, x_3, \ldots , x_N)^T \) from a group of values with mean \( \mu = (\mu_1, \mu_2, \mu_3, \ldots , \mu_N)^T \) and covariance matrix S is defined as:

\[
D_M(x) = \sqrt{(x - \mu)^T S^{-1}(x - \mu)},
\]

Each \( x \) is an observation which needs to be computed the Mahalanobis Distance score from the reference sample. And the \( \mu \) is the mean of the specific reference sample while covariance \( S \) is the covariance of the data in reference sample. According to the algorithm of Mahalanobis Distance, the quantity of instances in reference sample must be greater than the quantity of variate, namely, the dimension.

For multivariate normally distributed data the values are approximately chi-square distributed with \( p \) degrees of freedom \( (x_p^2) \). Multivariate outliers can now simply be defined as observations having a large (squared) Mahalanobis distance. For this purpose, a quantile of the chi-squared distribution (e.g., the 97.5% quantile) could be considered. After calculating the Mahalanobis Distance for a multivariate instance from the specific data group, we will get a squared Mahalanobis Distance score. If this score exceeds a “critical value”, this instance will be considered an outlier. The critical chi-square values for 2 to 10 degrees of freedom (equal to the number of variables under investigation) with corresponding probability level (alpha value) are shown below. When \( p < 0.05 \) we generally refer to this as a significant difference.

For example, the critical value for a bivariate relationship is 13.82. Any Mahalanobis Distances score above that critical value is a bivariate outlier. In our experiment, we will use the Mahalanobis Distance to find the outliers.

2.1 Workflow for Global Analysis using MD

For the global analysis, we calculate the Mahalanobis Distance for each instance from the whole dataset. This method is somewhat like the traditional measurement. The following diagrams indicate the Global Analysis mechanism for calculating the Mahalanobis Distance of the \( i^{th} \) and its next instance. The operation of global analysis using MD is visualized in Figure 1a.

2.2 Workflow for Cumulative Analysis using MD

With regard to the cumulative analysis, in our experiment, initially we calculate the Mahalanobis Distance of the first 50 records respectively. After that, for the \( i^{th} \) record, we treat the top \( i \) instances as the reference sample. The following diagrams indicate the Cumulative Analysis
mechanism for calculating the Mahalanobis Distance of the \(i^{th}\) and its next instance. The operation of cumulative analysis using MD is visualized in Figure 1(b).

2.3 Workflow for Lightweight Analysis with Sliding Window using MD

As to the lightweight analysis with sliding window, we propose a novel notion which called “sliding window”. The sliding window has a fixed size of a certain number of instances; and it moves forward to next instance when we analyze a new record. In our experiment, we set the window size to 50, 100, 200, 300, 500 and 1000. Initially we calculate each record’s Mahalanobis Distance in the window, respectively. After that, the window slides to the next record, with the corresponding size. For example, if we choose the window size of 50, each record within it will be computed the Mahalanobis Distance from the reference sample (namely, the selected 1 to 50 instances). Then, the window will slide forward by a step of one record. So, the window is formed with the instances of 2 to 51. That is to say, we calculate the 51st instance’s Mahalanobis Distance from the reference sample formed by records from 2 to 51. Next, it is supposed to compute the 52nd instance with a reference sample from 3 to 52, and so on. The following diagrams indicate the method about Lightweight Analysis with window size of 50 for calculating the Mahalanobis Distance of the \(i^{th}\) and its next instance. The operation of lightweight analysis using MD is visualized in Figure 1(c).

3. Methodology of Outlier Detection using Local Outlier Factor

3.1 Concept of Local Outlier Factor

Outlier ranking is a well-studied research topic. Breunig et al. (2000) have developed the local outlier factor (LOF) system that is usually considered a state-of-the-art outlier ranking method. The main idea of this system is to try to obtain an outlyingness score for each case by estimating its degree of isolation with respect to its local neighborhood. The method is based on the notion of the local density of the observations. Cases in regions with very low density are considered outliers. The estimates of the density are obtained using the distances between cases. The authors defined a few concepts that drive the algorithm used to calculate the outlyingness score of each point. These are the (1) concept of core distance of a point \(p\), which is defined as its distance to its \(k^{th}\) nearest neighbor, (2) concept of reachability distance between the case \(p1\) and \(p2\), which is given by the maximum of the core distance of \(p1\) and the distance between both cases, and (3) local reachability distance of a point, which is inversely proportional to the average reachability distance of its \(k\) neighbors. The LOF of a case is calculated as a function of its local reachability distance. In addition, there are two parameters that define the notion of density: a parameter MinPts specifying a minimum number of objects and a parameter specifying a volume. These two parameters determine a density threshold for the clustering algorithms to operate. That is, objects or regions are connected if their neighborhood densities exceed the given density threshold.

In [11], the author summarized the definition of Local Outlier Factor as follows:

Let \(D\) be a database. Let \(p, q, o\) be some objects in \(D\). Let \(k\) be a positive integer. We use \(d(p, q)\) to denote the Euclidean distance between objects \(p\) and \(q\).

Definition 1. (k-distance of \(p\))
The \(k\)-distance of \(p\), denoted as \(k\)-distance\((p)\) is defined as the distance \(d(p; o)\) between \(p\) and \(o\) such that:

(i) for at least \(k\) objects \(o' \in D \setminus \{p\}\) it holds that \(d(p, o') \leq d(p, o)\), and

(ii) for at most \(k-1\) objects \(o' \in D \setminus \{p\}\) it holds that \(d(p, o') < d(p, o)\).

Intuitively, \(k\)-distance\((p)\) provides a measure on the sparsity or density around the object \(p\). When the \(k\)-distance of \(p\) is small, it means that the area around \(p\) is dense and vice versa.

Definition 2. (k-distance neighborhood of \(p\))
The \(k\)-distance neighborhood of \(p\) contains every object whose distance from \(p\) is not greater than the \(k\)-distance, is denoted as
\[ N_k(p) = \{ q \in D(p) \mid d(p, q) \leq k\text{-distance}(p) \}. \]

Note that since there may be more than \( k \) objects within \( k\text{-distance}(p) \), the number of objects in \( N_k(p) \) may be more than \( k \). Later on, the definition of LOF is introduced, and its value is strongly influenced by the \( k\text{-distance} \) of the objects in its \( k\text{-distance} \) neighborhood.

**Definition 3.** (reachability distance of \( p \) w.r.t object \( o \))
The reachability distance of object \( p \) with respect to object \( o \) is defined as
\[ \text{reach-dist}(p, o) = \max \{ k\text{-distance}(o), d(p, o) \}. \]

**Definition 4.** (local reachability density of \( p \))
The local reachability density of an object \( p \) is the inverse of the average reachability distance from the \( k\text{-nearest-neighbors} \) of \( p \).
\[ lrd_k(p) = \frac{1}{\sum_{o \in N_k(p)} \text{reach-dist}(p, o)} \]

Essentially, the local reachability density of an object \( p \) is an estimation of the density at point \( p \) by analyzing the \( k\text{-distance} \) of the objects in \( N_k(p) \). The local reachability density of \( p \) is just the reciprocal of the average distance between \( p \) and the objects in its \( k\text{-neighborhood} \). Based on local reachability density, the local outlier factor can be defined as follows.

**Definition 5.** (local outlier factor of \( p \))
\[ \text{LOF}_k(p) = \frac{\sum_{o \in N_k(p)} \text{reach-dist}(p, o)}{|N_k(p)|} \]

LOF is the average of the ratios of the local reachability density of \( p \) and those of \( p \)'s \( k\text{-nearest-neighbors} \). Intuitively, \( p \)'s local outlier factor will be very high if its local reachability density is much lower than those of its neighbors.

In our experiment, we set the inspection effort to 0.1 which means that we regard the top 10% records as outliers according to the outlier score in decreasing sequence.

### 3.2 Workflow for Global Analysis using LOF

For the global analysis, we calculate the LOF score for each instance from the whole dataset. The following diagrams indicate the Global Analysis mechanism for calculating the LOF score of the \( i^{th} \) and its next instance. The operation of global analysis using LOF is visualized in Figure 2(a).

### 3.3 Workflow for Cumulative Analysis using LOF

As to the cumulative analysis, in our experiment, at first we calculate the LOF scores of the first 50 records respectively and labeled top 10% of the highest score ones as outliers. After that, for the \( i^{th} \) record, calculate the LOF score for all these \( i \) records, and then examine this \( i^{th} \) one to see whether it is among the top 10% highest score of the present dataset. If yes, then this instance is regard as an outlier. Otherwise, it is normal. The following diagrams indicate the Cumulative Analysis mechanism for calculating the LOF score of the \( i^{th} \) and its next instance. The operation of cumulative analysis using LOF is visualized in Figure 2(b).

### 3.4 Workflow for Lightweight Analysis with Sliding Window using LOF

The mechanism of lightweight analysis with LOF method to detect outliers is similar to the mechanism of Mahalanobis Distance method which mentioned above. The following diagram in Figure 2(c) indicates the method about Lightweight Analysis using LOF with window size of 50 to estimate an outlier of the \( i^{th} \) and its next instance.

**Figure 2.** Illustrations of outlier detection that can be done by using LOF in three different operation modes.

### 4. Experiment for Comparison

In this paper, we use a dataset named “household power consumption” from MOA Massive Online Analysis for our experiment. We choose 10,000 instances for analysis in our experiment. As a traditional distance-based method, Euclidean distance should not be omitted. So we also try this classic method in our experiment and use its performance as a benchmark. In the following, we briefly review the basics of estimation theory about the Euclidean distance.

\[ D_{\text{Euclidean}}(x_i) = \sqrt{(x_i - \mu_i)^T(x_i - \mu_i)} \]

where \( \mu_i \) represents the mean vector of class \( i \) and \( x_i \) represents the sample vector to classify.
We use Global Analysis and Cumulative Analysis as the benchmark respectively and make the comparison with Lightweight Analysis. True Positive Rate (TPR) and False Positive Rate (FPR) are used as the metric. For a comprehensive analysis, we use two standards (hard standard and soft standard) to conduct the experiment with all the methods mentioned above. For the Euclidean distance and Mahalanobis distance, the hard standard is set with the probability of 0.001 and 0.05 of the soft one. With regard to the LOF method, the minPtsLB and minPtsUB are set to 10 and 40 respectively for hard standard and 50, 80 for soft standard. The following bar charts indicate the result.

In fact, the Cumulative Analysis is closer to the real situation. Imagine that we use the traditional analysis method to detect the outlier. What we need is the whole dataset we got at this very moment, like 1000 instances. Then we conduct the analysis for these 1000 records and get the result. As our paper’s assumption, our object is data stream which comes to the database continuously. That means we need to analyze the data continuously. Suppose that after a short while, there are another 200 new datasets coming into the database. For the traditional analysis, we should use all the dataset we currently have, which up to 1200 records. So we do the outlier detection for these 1200 instances. The Cumulative Analysis in our experiment faithfully simulates this situation. But we still implement the Global Analysis as a reference.

The novel method “Lightweight Analysis with Sliding Window” we propose in this paper is real-time, flexible and efficient. What we need to do is set a window size and the scales for the window to slide with based on actual situation. And the analysis is aimed at the data within the window. Just like the workflow we mentioned above.

4.1 Result of using Euclidean Distance

We want to test the efficacy of Lightweight Analysis with effects of different window sizes, versus Global Analysis and Cumulative Analysis. This represents a test of outlier detection that is done on the fly (incremental) against using most of the static data points – Global Analysis uses all, and Cumulative Analysis uses all the accumulated data points that have been received up to the current processing time so far.

The results in Figures 3(a) and 3(b) are produced by Lightweight Analysis and they are the accuracies derived from the proportion of outliers being discovered in relative to the benchmarks of Global Analysis (as in Figure 3(a)) and Cumulative Analysis (as in Figure 3(b)) respectively. In each of these Figures, two levels of outlier detections are being used – hard standard and soft standard. From the Figures, we can clearly see that under the hard standard, the TPR is higher than that of soft standard. It is because in the hard standard mode, the Euclidean distance method can only find very few outliers. Once we relax the condition, it detects a lot more outliers, but the accuracy deteriorates considerably.
4.2 Result of using Mahalanobis Distance

The same experiments are repeated but by using Mahalanobis Distance (MD). As we can see in Figures 6(a) and 6(b), the larger the window size, the higher accuracy rate we can get. If we use a looser standard, we not only find more outliers but we also have a slight improvement in the accuracy, which is certainly an advantage of MD over the traditional Euclidean distance method.

![Figure 6(a). Accuracy of outlier detection by Lightweight Analysis using Mahalanobis Distance, wrt Global Analysis.](image)

![Figure 6(b). Accuracy of outlier detection by Lightweight Analysis using Mahalanobis Distance, wrt Cumulative Analysis.](image)

More importantly, our proposed Lightweight Analysis method can detect more outliers than the Cumulative Analysis. The reason is that the Lightweight Analysis method focuses on the datasets within the window, which is more sensitive to some regional outliers. However the Cumulative Analysis concentrates on all the current datasets, which evaluates outliers based on the overall average. So our proposed Lightweight Analysis makes a significant improvement against the traditional one.

The number of outliers found by different methods of measurement using MD are shown in Figures 7(a) and 7(b), in hard standard and soft standard respectively as follows.

![Figure 7(a). Number of Outliers Found in Euclidean Dist. Analysis with Hard Standard.](image)

![Figure 7(b). Number of Outliers Found in Euclidean Dist. Analysis with Soft Standard.](image)
4.2 Result of using Local Outlier Factor

Again the experiments are repeated by using Local Outlier Factor (LOF) as a distance measurement method. From the comparison charts shown in Figures 7(a) and 7(b), we can see that the window size is no longer the bigger the better. Specifically, after the window size of 300, the accuracy rate is trending downward. So, we may infer that the LOF method has an appropriate or optimal window size. Moreover this method has a higher false positive rate than the Mahalanobis Distance method.

The 3-D plots in Figures 8 (a)-(d) visualize the data with color map for cumulative analysis and lightweight analysis with window size equals to 1000. The purpose is to visually compare how Lightweight Analysis performs wrt to Cumulative Analysis using LOF, because the TPR of Lightweight Analysis wrt to Cumulative Analysis is unusually low when window size is large at 1000, as shown in Figure 7(b).
Figure 8(c). Visualization of outliers detected by soft standard of Cumulative Analysis using LOF.

Figure 8(d). Visualization of outliers detected by soft standard of Lightweight Analysis using LOF.

From the above visualization plots we can see that the Lightweight Analysis has much better performance than the Cumulative Analysis. The Lightweight Analysis method detects a lot more outliers which seem to be real outliers in the dataset. Now we can explain why the TPR in Lightweight Analysis against Cumulative Analysis is usually low in window 1000. Because the latter one did not thoroughly find the outliers but the former one did. In other words, many outliers found in Lightweight Analysis are regarded as normal in Cumulative Analysis thus, leads to a low TPR. This experiment demonstrates that for the LOF method, our proposed Lightweight Analysis performs somewhat better than the traditional one.

Lastly the numbers of outliers found by different methods of measurement are shown in Figures 9(a) for hard standard and 9(b) for soft standards. In both cases, it can be observed that in the context of LOF measurement, the Lightweight Analysis with small window size method is largely producing more outliers. It can find as many outliers as Global Analysis does, after processing about 6000 data points. Lightweight Analysis with small window size started to outperform Cumulative Analysis soon after processing 5900 data points in both hard and soft standards. Lightweight Analysis indeed can find more outliers with small window than large ones using LOF.

A similar phenomenon is observed (c.f. Figure 6(b)) in MD but only in soft standard. In contrast, Lightweight Analysis with small window finds only very few outliers in Euclidean Distance.

5. Conclusion

This paper proposed a general framework for finding outliers in an incremental fashion. A collection of methods in different operation modes and distance measurements are formulated and experimented. It contributes to outlier detection in data stream mining which is important but a relatively neglected research area.

Acknowledgement

The authors are thankful for the financial support from the research grant of Grant no. MYRG152(Y3-L2)-FST11-ZY, offered by the University of Macau, RDAO.

References


