# Performance Evaluation of Ensemble Method Based Outlier Detection Algorithm

Priya. M[1], M. Karthikeyan[2]
*Department of Computer and Information Science, Annamalai University,*
*Annamalai Nagar, Tamil Nadu, India [1,2]*
*Email:mpriyaau@gmail.com[1], karthiaucse@gmail.com[2]*

**Abstract-** Outlier analysis is a necessary research in data mining. Outlier detection can be used in many application areas like the diagnosis of diseases, fraud detection, agricultural, etc. so that there is a need to detect Outliers. In the proposed method, we focused to analyze the performance of outlier detection algorithm using feature bagging technique on health care application. The density-based Local Outlier Factor algorithm is used to detection of outliers. The local density of an object depends on its K nearest neighbor of objects. Next we introduce feature bagging technique which is one of the ensemble methods. Ensemble method of classifiers has been effective in improving overall performance and stability of machine learning algorithms. Local Outlier Factor method and LOF with feature bagging technique observed that the performance of the local outlier factor algorithm with feature bagging technique improves the accuracy based n diabetic dataset.

**Keywords-**Outliers analysis; Data Mining; Outlier Detection; Bagging; Local Outlier Factor.

## 1. INTRODUCTION

The growth of big data research in recent years, data mining and outlier detection techniques are growing fast. Data mining deals within the process of the discovered non-trivial, hidden and interesting knowledge from data. Recently, Detection of outliers has gained a lot of attention in several domains, such as detecting fraudulent transactions, medical diagnostics and intrusion detection to direct marketing.

In the outlier detection techniques, we can detect the data objects which are significantly different from the other data objects. Outliers are infrequent data objects, in each object are compared to other objects, making their detection is very extremely important. The outliers detected from database to have a fraud behavior. In Data mining, problems are solved based on both supervised learning and unsupervised learning. Supervised learning techniques build a prediction model for data objects based on the training set (labeled data), and used to classify each data object. Unsupervised learning techniques does not require training set labeled data and detect outliers as data objects which are different from the normal data object. These techniques are called outlier detection techniques. Outlier detection algorithms can detect rare objects as deviations from normal behavior.

By definition, an outlier is “an observation of the data that deviates from other observations so much that it arouses suspicions that it was generated by a different and unusual mechanism” [1]. Outliers are erroneous or real. Sometimes outliers can be found in an output of clustering techniques. The clustering techniques defined the outliers are objects, that does not lie in any of the groups formed. Thus, the clustering methods defined as outliers are the noise in which the clusters can be embedded. Another technique defines outliers as objects, that are neither a group of a cluster nor a part of the noise; but outliers are objects which behave differently from the normal data. Outliers are sometimes often being considered as a single object or group of objects that exhibits behavior outside the normal range.

Outlier detection is engaged to measure the distance between data objects to detect those objects which are totally different from or inconsistent with the remaining set of data [2]. In data mining, outlier detection problem is one of the most fundamental issues. The contribution in this paper combines local outlier factor algorithm with feature bagging technique and also application of feature bagging to diabetic data set has not been explored so far. This led to the motivation to analyze the performance of Feature bagging technique in diabetic dataset.

This paper is organized as follows. The related work is presented in section 2. Outlier detection technique and the local outlier factor algorithm discussed in section 3. The experimental results are provided in section 4 and a section 5 concludes the paper.

## 2. RELATED WORK

Outlier detection in data mining has been a number of real life applications in areas such as medical care diagnosis, fraud detection, and identification and emerging business trends in e-commerce.

Priya and Karthikeyan have discussed comparison of different clustering algorithms for outlier detection. The comparison has been made to detect outliers in the health care datasets by using clustering algorithms [3]. Agyemang et al. focused the survey of practical applications of outlier mining, and provides classification for categorizing related mining techniques [4]. Recently many algorithms use the proximity concepts in order to find outliers based on their relationship to the rest of the data. However, in high dimensional space, the data is sparse and the notion of proximity fails to retain its meaningfulness. In fact, high dimensional data implies that every object is an almost equally good outlier from the perspective of proximity-based definitions [5]. Consequently, for high dimensional data, the notion of finding meaningful outliers becomes substantially more complex and nonobvious.

Breuning et al. proposed LOF method based on the local density of given data objects locality to identify local outlier. This method detects outlier by comparing the object's local density to its neighbor local density and computes the local outlier factor (LOF) for each object. The LOF is used to identify outliers that different from their cluster. The LOF may not be efficient in density with mere neighbors and fails to identify outlier when neighbors have analogous compactness [6].

Gupta et al. have proposed an overview of the various techniques for outlier detection on temporal data. Modeling temporal data is a challenging task due to the dynamic nature and complex evolutionary patterns in the data. They have discussed with different data types, presented in various outlier detections[7].

Lazarevic and Kumar have focused traditional ensemble approach for detecting outliers in high dimensional and noisy databases [8]. The outlier detection algorithm uses a small set of features which are randomly selected from the original feature set. As a result, each outlier detector identifies different outliers, and thus assigns to all data records outlier scores that correspond to their probability of being outliers.

Minh Quoc et al. have proposed randomized algorithm in which one can compute local outlier factor very efficiently for high dimensional datasets using random points [9].

## 3. METHODOLOGY

Outlier detection is a necessary task in data mining. Outlier detection has several important application areas and that merits ever more consideration from data mining community. In this paper, outlier detection algorithm is proposed by computing the distances of the objects from one another as well as on computing the densities of local neighborhoods. The dense region of objects in the data space considered as clusters in the Density based method. These dense regions are separated from low density regions in which represent outliers (noise).

### *3.1. Density Based Local Outlier Factor (LOF) Detection Algorithm*

Density based techniques estimate the density of an object x within a small region by counting number of objects within a neighborhood region. An object x of local density depends on its K nearest neighbor objects. The local outlier detection algorithm assigns to each object with a degree can be an outlier. This degree is known as the local outlier factor (LOF) of an object. The degree depends on it is local in that; how it has isolated the object is with respect to encompass its neighborhood.

The LOF value of all data objects are sorted in decreasing order. The high LOF value of data objects are outliers whereas the low LOF values of data objects are possible to be normal objects with regard to their neighborhood. High LOF value indicates the low density neighborhood and therefore high capability of being an exception.

Figure 1 shows simple example of the LOF approach based on nearest neighbor method. Let us consider a two-dimensional data set; it is evident that the density of the C2 cluster is significantly more than the density of the C1 cluster. The low density of the cluster C1 for object p3 is inside the cluster C1. The distance between the object p3 and its nearest neighbor is most similar to the distance between the object p2 and the nearest neighbor from the cluster C2, but the object p2 is not be considered as outlier in the nearest neighbor (NN) method. In another way, LOF approach is capable to capture the object p2 as outlier as a result of the fact that it considers the density around the objects. Nonetheless, the object p1 may be identified as outlier by using both NN and LOF approaches, since it is far away from both clusters.

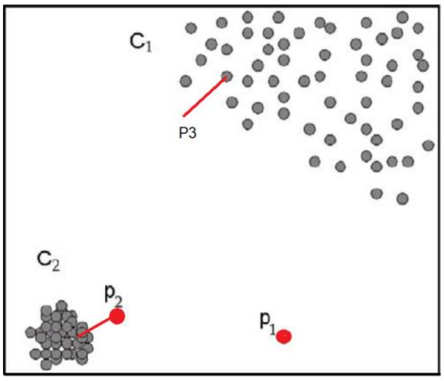

**Figure 1 The LOF approach**

3.1.1. *LOF algorithm*

Let D be a dataset, p1, p2, p3 are some objects in D and k be a positive integer. The distance between objects p1 and p2 denoted as $d(p1, p2)$, that is the Euclidean distance function.

Step 1: Calculate k- distance of (p1) that provides a measure of the density around the object p1, when k-distance of p1 is less that the area surrounded by p1 is dense and vice versa.

Step 2: Finding k-distance neighborhood of (p1): $\mathrm{dist}_k(p1) -$ the k-distance neighborhood of p1 contains every object whose distance for p1 is not greater than the k-distance.

Step 3: Calculate reachability distance of p1 with respect to object p3 as $\mathrm{Reachdist}_k\,(p1, p3) = \max\{\mathrm{dist}_k(p3), d(p1, p3)\}$

Step 4: Compute the local reachability density of p1. The local reachability density of an object p1 is the inverse of the average reachability distance from the k-nearest neighbors of p1.

Step 5: Finally find Local outlier factor of p1.
The local outlier factor is a ratio that determines whether or not an object is outlier with respect to its neighborhood.

LOF (p1) is the average of the ratios of the local reachability density of p and that of p's k-nearest-neighbors.

Computing reachability distance of object p1 involves the process of computing distances of all objects within the neighborhood of p1. This distance is compared with the k-distance of that neighborhood. But this approach is very expensive. Secondly, for every object LOF computation is to be done before the few outliers are detected.

### *3.2. Feature Bagging*

Ensemble method helps to improve machine learning outputs by combining various models. This method allows the result of better predictive performance compared into a single model. They are meta-algorithms which combines many machine learning algorithms into one predictive model in order to improve predictions. Common types of ensembles are Bootstrap aggregating, Boosting, Stacking, Bayes optimal classifier etc.

Bagging stands for bootstrap aggregation. It is an ensemble meta-algorithm described to improve the accuracy and stability of machine learning techniques. Feature bagging is one of the random subspace methods, which attempts to reduce the correlation between estimators in an ensemble by training them on random samples of features instead of the entire feature set.

3.2.1. *Feature bagging algorithm*

The feature bagging algorithm is explained in the following steps

```
Step 1: Normalize the data set D and the data set has d
        as dimension.
Step 2: Repeat steps (i) to (iii)
        (i) Choose randomly variable size in a subset n.
        (ii) Select randomly n variables without
             replacement.
        (iii) Apply LOF algorithm to the subset data.
Step 3: Combine the outlierness scores and find the
        objects of outliers.
```

## 4. EXPERIMENTAL RESULTS

In machine learning most of the research work related to the application area of health care diagnosis has focused on the data set in the UCI repository. In order to study the efficiency of this method, we have used diabetic data set is explained in the following.

### *4.1. Dataset description*

Diabetic data set contains the 768 Dimensions of data records (objects). Each record contains 8 attributes which are considered as factors for the occurrence of diabetic, like (1) Number of times pregnant (2) Plasma glucose concentration a 2 hours in an oral glucose tolerance test (3) Diastolic blood pressure(mm Hg) (4) Triceps skin fold thickness (mm) (5) 2-Hour serum insulin (mu U/ml) (6) Body mass index (weight in kg/(height in m)^2) (7) Diabetes pedigree function (8) Age (years) (9) Class variable (0 or 1). The output class variable labeled as 0 or 1(class value 1 is for diabetes and 0 is for non-diabetes). With these attributes we try to classify whether the presence or absence of diabetic using the newly developed model in this paper.

### *4.2. Performance Evaluation*

The F1 score can be explained as a harmonic average of precision and recall. It is a measure of a test's accuracy. It considers both the precision p and the recall r of the test to compute the score. Precision p is the number of correct positive results divided by the number of all positive results. Recall r is the number of correct positive results divided by the number of positive results that should have been returned.

The AUC (Area Under the Curve) is used to measure the outlier detection algorithm performance. The AUC is defined as the surface area under its Receiver Operating Characteristic (ROC) curve. The AUC for the perfect ROC curve is set to be 1, while AUCs of "less than perfect" outlier detection algorithms are less than 1. It is one of the most important evaluation metrics for checking any model's performance [10].

The AUC curve is designed by plotting the true positive rate (TPR) versus the false positive rate (FPR) at different threshold value settings. The true-positive rate is also called as sensitivity or recall (probability of detection) in machine learning. The false-positive rate is called as the fall-out (probability of false alarm) and it can be computed as (1 - specificity). So the ROC curve is the sensitivity as a function of fall-out. In general, if the probability distributions for both detection and false alarm are known, the ROC curve can be generated by plotting the cumulative distribution function area under the probability distribution from $(-\infty)$ to the discrimination threshold of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability in x-axis [11].

### *4.3. Analysis of Results*

Initially, LOF algorithm was applied to diabetic dataset for detecting outliers. After that, Feature bagging was applied to the output of local outlier factor algorithm. The k values were varied from 3 to 9. Compute F1 Score, R-precision, Average precision and ROC-AUC for various k values, as shown in table 1. From the table 1, comparison of different evaluation measures, it is observed that feature bagging technique is performed well than that of the local outlier factor algorithm.

Table 1. Evaluation Measure.

| **K** | **F1 score** | | **R-precision** | | **Average precision** | | **ROC AUC** | |
|---|---|---|---|---|---|---|---|---|
| value | LOF | LOF + FB | LOF | LOF + FB | LOF | LOF + FB | LOF | LOF + FB |
| 3 | 0.37 | 0.38 | 0.20 | 0.25 | 0.21 | 0.31 | 0.50 | 0.58 |
| 6 | 0.36 | 0.37 | 0.20 | 0.25 | 0.20 | 0.26 | 0.47 | 0.53 |
| 9 | 0.35 | 0.39 | 0.21 | 0.21 | 0.20 | 0.24 | 0.48 | 0.56 |

The evaluation measures, F1 score, R-precision, Average precision and ROC AUC variations are represented in figure 2, 3, 4 and 5 respectively. From the figure, there is some improvement in the result when feature bagging technique is applied.

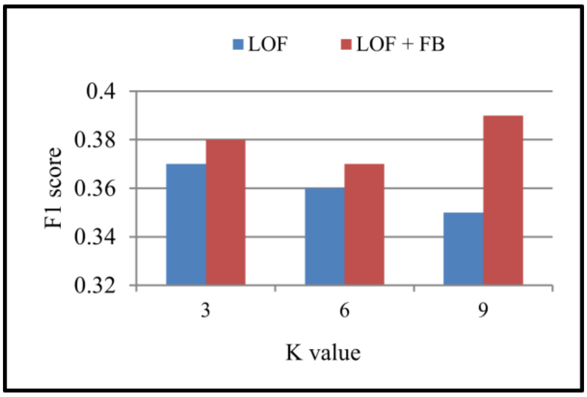

Figure 2 F1 Score

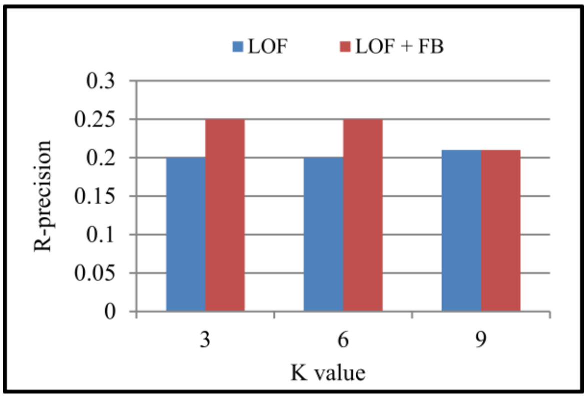

Figure 3 R-Precision

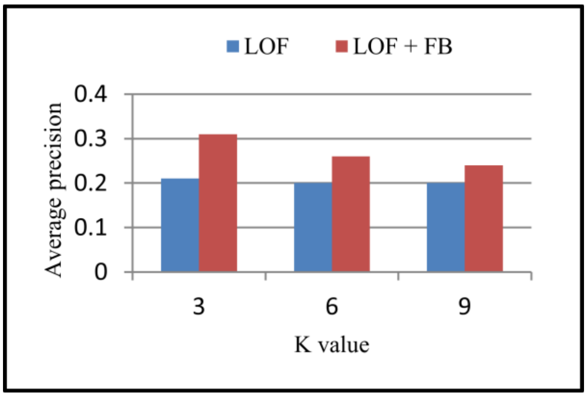

Figure 4 Average precision

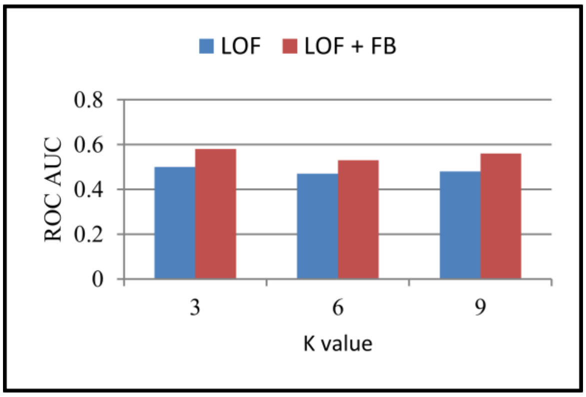

Figure 5 ROC AUC

## 5. CONCLUSION

In this paper, local outlier factor detection algorithm is employed on diabetic experimental dataset then local outlier factor algorithm with feature bagging is also employed to detect the presence of outliers. Performance evaluation analysis is measured using the F1 score, R-Precision, Average precision, and ROC-AUC measures. The result shows that performance on varying k value has improved to identify the outliers when we adopt the feature bagging technique. More detailed study has to be carried out on this aspect. However, this method can be extended to detection of outliers for different types of dataset.

Detection of outliers in health care applications in data mining has received much attention is more powerful tool, which provides an accurate extraction. This method provides an alternate methodology to detect the diabetic people in medical or clinical side to provide a consistent, better and efficient health care services.

## REFERENCES

[1] D.M. Hawkins, Identification of Outliers, Chapman and Hall, London, 1980.

[2] J.Han, & Kamber, M. (2006), Data Mining: Concepts and Techniques, Second edition, Morgan Kaufmann Publishers, pp. 285–464.

[3] M. Priya and M. Karthikeyan, (2018) "A Comparative Study Of Clustering Algorithms For Outlier Identification", International Journal for Research in Engineering Application & Management (IJREAM), Vol-04, Issue-07, pp. 391-396. ISSN : 2454-9150

[4] M. Agyemang, Barker, K., & Alhajj, R. (2006), "A comprehensive survey of numeric and symbolic outlier mining techniques", Intelligent Data Analysis 10 (6) 521–538.

[5] C. Aggarwal and P. Yu, (2001), "Outlier Detection for High Dimensional Data". In Proceedings of the ACM SIGMOD International Conference on Management of Data, Volume 30, Issue 2, pages 37 – 46.

[6] M. M. Breunig; Kriegel, H.-P.; Ng, R. T.; Sander, J. (2000), "LOF: Identifying Density-based Local Outliers". Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data. SIGMOD: 93–104.

[7] Gupta, Manish, et al, (2014), "Outlier detection for temporal data." Synthesis Lectures on Data Mining and Knowledge Discovery 5.1: 1-129.

[8] A.Lazarevic, Kumar, V., (2005), "Feature bagging for outlier detection". Proc. 11th ACM SIGKDD international conference on Knowledge Discovery in Data Mining: 157–166.

[9] Nguyen, Minh Quoc, et al., (2010), "A fast randomized method for local density-based outlier detection in high dimensional data." Data Warehousing and Knowledge Discovery. Springer Berlin Heidelberg.pp. 215-226.

[10] P.Andrew and Bradley, (1997)," The use of the area under the ROC curve in the evaluation of machine learning algorithms." Pattern Recognition Vol. 30, Issue 7, pp. 1145-1159.

[11] https://en.wikipedia.org/wiki/Receiver_operating_characteristic