I J C T A, 9(2) 2016, pp. 591-596

Performance Evaluation of Ensemble Based Outlier Detection Method for Health Care Application

V. Mahalakshmi[1] and M. Govindarajan[1]

ABSTRACT

Outlier detection is an important research area that forms part of numerous application areas. Outlier detection is used in different applications like intrusion detection, fraud detection, track environmental changes, medical diagnosis so there is need to detect outliers. Various approaches are used for outlier detection. In this research paper, we focused outlier detection technique on cardiovascular data set available in R package. We analyze the performance of feature bagging technique for the used data set in terms of F1 score, Average precision, R-precision and ROC-AUC. The performance of Local Outlier Factor algorithm was found to improve on employing feature bagging technique.

Keywords: Data mining, Outlier Detection, Bagging, Density-based method, Local Outlier Factor.

I. INTRODUCTION

Data mining, in general, deals with the discovery of hidden, non-trivial and interesting knowledge from different types of data. As the advancement of information technologies is taking place, the quantity of databases, as well as their dimension and complexity is developing rapidly. Hence there is a requirement for automated analysis of great amount of information. The analysis results which are then generated are utilized for making a decision by a human or a program. Outlier detection is one of the basic problems of data mining.

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 [6]. Outliers may be erroneous or real. Outliers are sometimes found as a side-product of clustering techniques. These techniques define outliers as points, which do not lie in any of the groups formed. Thus, the procedures implicitly define outliers as the background noise in which the clusters are embedded. Another class of methods is also there which defines outliers as points, which are neither a part of a cluster nor a part of the background noise; but specifically are points which behave very differently from the normal data. Sometimes outliers can often be considered as a single individual or group of individuals that exhibits behavior outside the normal range.

Outlier detection is employed to measure the distance between data objects to detect those objects that are entirely different from or inconsistent with the remaining data set [6]. Data that appear to have different characteristics than the rest of the population are called outliers [1]. The problem of anomaly detection is one of the most fundamental issues in data mining. The contribution in this research work combines local outlier factor algorithm with feature bagging technique. Also application of feature bagging to heart data set has not been investigated so far. This led to the motivation to analyse the performance of Feature bagging technique in cardiovascular dataset.

* Assistant Professor, Department of Computer Science and Engineering, Faculty of Engineering and Technology, Annamalai University, Annamalai Nagar, India, *E-mail: mahaa80@gmail.com, govind_aucse@yahoo.com*

The organization of this paper is as follows. The related work is presented in section 2. Outlier detection technique is discussed in section 3 and the local outlier factor algorithm introduction is given here. The experimental results are provided in section 4 and a section of concluding remarks follows.

RELATED WORK

Outlier mining has been akin to finding needles in a haystack. However, outlier mining has a number of practical applications in areas such as fraud detection, network intrusion detection, and identification of competitor and emerging business trends in e-commerce. In [1] the survey discuses practical applications of outlier mining, and provides a taxonomy for categorizing related mining techniques.

K. Breuning *et al.* [2], proposed LOF Technique based on the local density of given trial's locality to identify local outlier. A local outlier factor (LOF) is computed for each point. The LOF mines outliers that deviate from their cluster. The LOF may not be efficient in density with meagre neighbors and fails to calculate outlier when neighbors have analogous compactness.

Many recent algorithms use concepts of proximity 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, the scarcity of high dimensional data implies that every point is an almost equally good outlier from the perspective of proximity-based definitions [3]. Consequently, for high dimensional data, the notion of finding meaningful outliers becomes substantially more complex and non-obvious.

In the paper [4] focus is given on establishing the context of recently proposed replicator neural network (RNN) approach for outlier detection. This approach employs multi-layer perceptron neural networks with three hidden layers and the same number of output neurons and input neurons to model the data. The neural network input variables are also the output variables so that the RNN forms an implicit, compressed model of the data during training. A measure of outlyingness of individuals is then developed as the reconstruction error of individual data points.

In the work [5], an organized overview of the various techniques for outlier detection on temporal data is proposed. Modeling temporal data is a challenging task due to the dynamic nature and complex evolutionary patterns in the data. This survey organized the discussion along different data types, presented various outlier deûnitions, and brieûy introduced the corresponding techniques.

Outlier ensemble based methods are investigated in [7]. The methods may be considered traditional as they define outlier without regard to class membership.

Minh Quoc Nguyen *et al.* proposed randomized algorithm [10] which can compute local outlier factor very efficiently for high dimensional datasets in which random points has been selected which makes data partitioned which is based on consistency property of outliers.

Local Outlier Factor (LOF) an approach of new algorithm was used to detect an outlier from the input data. In [11], LOF algorithm proposes less computational time and also dynamically updating the profiles of data streams. This algorithm was also found to be computationally very efficient as compared to other algorithms.

METHODOLOGY

Outlier detection technique

Outlier detection is an important task in data mining. Outlier detection has many important applications and deserves more consideration from data mining community. It is an important branch in data mining, as this stage is required in elaboration and mining of information coming from many application fields such

as industrial procedures, transportation, biology, public safety, climatology [12]. Anomalies are information which can be considered peculiar due to several causes.

Researches on outlier detection broadly fall into following categories. Distribution based methods, Depth based algorithms, Deviation based techniques, Distance based algorithms, Density based methods, Sub Space based techniques, etc. Density based methods regard clusters as dense region of objects in the data space. These dense regions are separated by low density regions which represents noise.

A. LOF: Identifying Density-Based Local Outliers

Density based techniques measure density of a point x within a small region by counting number of points within a neighborhood region. Breunig et al. [2] introduced a concept of local outliers which are detected based on the local density of points. Local density of a point x depends on its K nearest neighbor points. The density based local outlier detection scheme assigns to each object a degree to be an outlier. This degree is called the local outlier factor (LOF) of an object [8]. It is local in that, the degree depends on how isolated the object is with respect to the surrounding neighborhood.

All data points are sorted in decreasing order of LOF value. Outliers are data objects with high LOF values whereas objects with low LOF values are likely to be normal with respect to their neighborhood. High LOF is an indication of low-density neighborhood and hence high capability of being an exception.

Top m points are chosen as outliers from this sorted order. However, LOF does not function admirably with large datasets as it computes a large number of distance calculations in order to ûnd nearest neighbors of each pattern in the dataset. The number of nearest neighbors is represented by ‘k’.

We briefly review the steps of computing LOF value of an object p in a dataset D:

Let D be a database, p, q, o some objects in D and k be a positive integer. The distance function (Euclidean distance) d (q, p) is used to denote the distance between objects p and q.

Step 1: Computing (k- distance of p): k-distance (p) provides a measure of the density around the object p, when k-distance of p is small meaning that the area around p is dense and vice versa.

Step 2: Finding (k-distance neighborhood of p): The k-distance neighborhood of p contains every object whose distance for p is not greater than the k-distance.

Step 3: Computing (reachability distance of p wrt object o): The reachability distance of object p with respect to object o is Reach-dist-k (p, o) = max {k- distance (o), d (p, o)}.

Step 4: Computing (the local reachability density of p): The local reachability density of an object p is the inverse of the average reachability distance from the k-nearest neighbors of p.

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

Step 5: Local outlier factor of p: The local outlier factor is a ratio that determines whether or not an object is outlier with respect to its neighborhood. LOF (p) is the average of the ratios of the local reachability density of p and that of p’s k-nearest-neighbors.

The major limitation of the LOF algorithm [1] lies in computing reachability distance which is defined as reach-dist-k (p, o) = max (k-distance (o), d (p, o)).

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

B. Feature Bagging

In statistics and machine learning, ensemble methods use various learning algorithms to get better predictive performance than could be obtained from any of the constituent learning algorithms. Unlike a statistical ensemble in statistical mechanics, which is usually infinite, a machine learning ensemble refers only to a concrete limited set of alternative models, but allows for much more adaptable structure to exist among those options. Common types of ensembles are Bayes optimal classifier, Bootstrap aggregating, Boosting, Stacking etc.

Bootstrap aggregating, also called Bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. Although the major application is done to decision tree methods, it can be used with any type of method. Bagging is a special case of the model averaging approach.

EXPERIMENTAL RESULTS

(A) Environment

ELKI is an open source (AGPLv3) data mining software written in Java aimed at users in research and algorithm development, with an emphasis on unsupervised methods such as cluster analysis and outlier detection. The ELKI visualization tools have been extended to support (the clustering of) uncertain data.

(B) Dataset description

The Cardiovascular dataset in R package was used for outlier detection. This database contains 13 attributes. The attributes are age, sex, chest pain type, resting blood pressure, serum cholesterol in mg/dl, fasting blood sugar, resting electrocardiographic results, maximum heart rate achieved, exercise induced angina, ST depression induced by exercise relative to rest, the slope of the peak exercise ST segment, number of major vessels, and thal. With these attributes, the absence or presence of heart disease can be predicted.

(C) Performance evaluation

The F1 score can be interpreted as a weighted average of precision and recall. 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 area under the curve (AUC) is used to measure the performance of outlier detection algorithm. The AUC of specific algorithm is defined as the surface area under its Receiver Operating Characteristic(ROC) curve. The AUC for the ideal ROC curve is typically set to be 1, while AUCs of “less than perfect” outlier detection algorithms are less than 1.

In statistics, a receiver operating characteristic (ROC), is a graphical plot that illustrates the performance of a binary classifier system on its discrimination threshold variation. The curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) by setting various threshold values. The true-positive rate is also known as sensitivity or the sensitivity index *d’*, known as “d-prime” in signal detection and biomedical informatics, or recall in machine learning. The false-positive rate is also known as the fall-out and can be calculated as (1 - specificity). The ROC curve is thus 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 (∞ to ∞) of the detection probability in the y-axis versus the cumulative distribution function of the false-alarm probability in x-axis.

(D) Analysis of results

Initially, LOF algorithm was applied to heart dataset for detecting outliers. Later Feature bagging was applied to the output of local outlier factor algorithm. Euclidean distance function was applied. The values of k were varied from 2 to 8. For various values of k, Maximum F1, R-precision, Average precision and ROC-AUC were computed as shown in table 4.1. On comparison the values of the evaluation measures seems to increase on applying feature bagging technique when compared to that of the local outlier factor algorithm output.

Table 4.1
Evaluation measures for varying k values

K value	*Maximum F1*		*R-precision*		*Average precision*		*ROC AUC*	
	LOF	*LOF + FB*	*LOF*	*LOF + FB*	*LOF*	*LOF + FB*	*LOF*	*LOF + FB*
2	0.37	0.38	0.21	0.25	0.24	0.26	0.52	0.58
4	0.36	0.39	0.17	0.25	0.21	0.31	0.50	0.58
6	0.35	0.35	0.20	0.31	0.20	0.26	0.47	0.53
8	0.35	0.40	0.20	0.20	0.20	0.24	0.48	0.56

The variation of the four evaluation measures, Maximum F1, R-precision, Average precision, and ROC AUC are represented in figures 4.1, 4.2, 4.3 and 4.4 respectively. From the chart it is incident that there is some improvement in the output when feature bagging technique is employed.

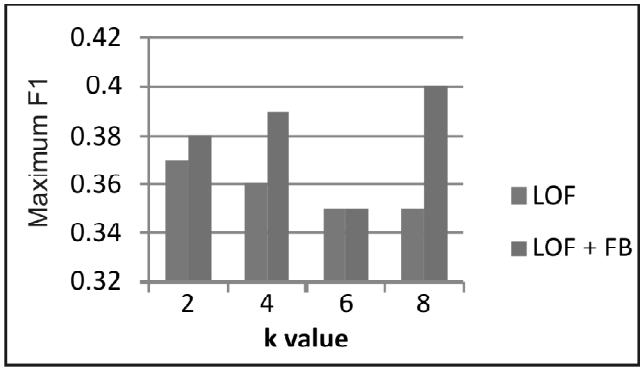

Figure 4.1: Maximum F1

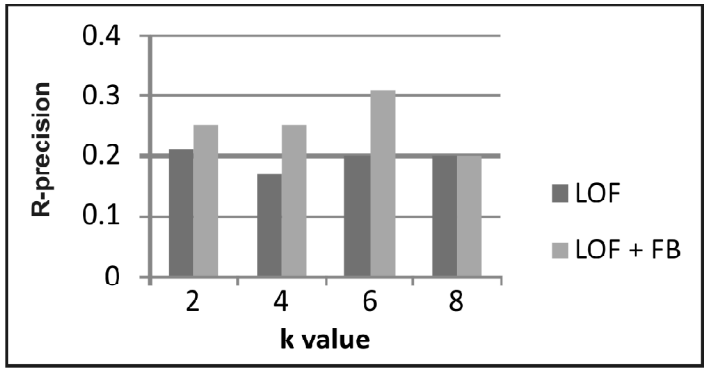

Figure 4.2: R-precision

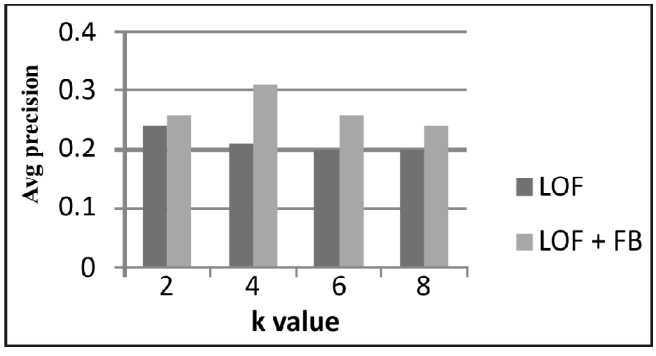

Figure 4.3: Average Precision

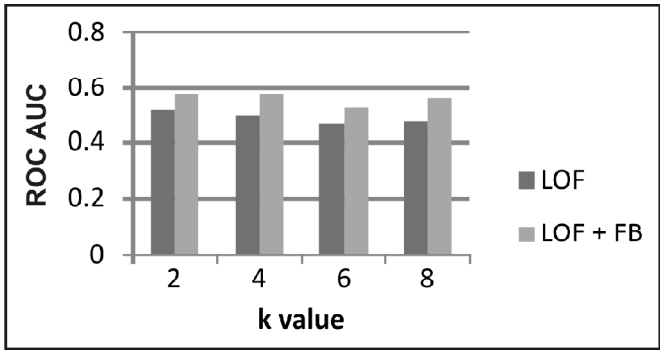

Figure 4.4: ROC AUC

CONCLUSION

Outlier detection as a branch of data mining has many important applications and deserves more attention from data mining community. In this research work, outlier detection algorithm local outlier factor is

examined. From the experimental analysis of the heart dataset, we can see that local outlier factor algorithm along with feature bagging can be used for outlier analysis. Performance analysis is measured by using the following measures, namely, Maximum F1 score, R-Precision, Average precision, and ROC-AUC. The work presented here shows improved performance on varying the values of k by employing feature bagging technique. However, the work can be enhanced in future to give better performance by employing various outlier detection algorithms. The same can be extended for various datasets also. Detection of outliers in medical data applying data mining technique is more powerful and provides accurate extraction results. The process helps the medical decision makers to provide better, consistent and efficient healthcare services.

REFERENCES

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

[2] Breunig, M. M.; 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.

[3] 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.

[4] G. Williams, R. Baxter, H. He, S. Hawkins and L. Gu, (2002), "A Comparative Study for RNN for Outlier Detection in Data Mining". In Proceedings of the 2nd IEEE International Conference on Data Mining, page 709, Maebashi City, Japan.

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

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

[7] Lazarevic, A.; Kumar, V., (2005), "Feature bagging for outlier detection". Proc. 11th ACM SIGKDD international conference on Knowledge Discovery in Data Mining: 157–166.doi:10.1145/1081870.1081891.

[8] Mansur, M. O., Md Sap, and Mohd Noor, (2005), "Outlier detection technique in data mining: a research perspective." pp. 23-31.

[9] Nelson Gnanaraj, Dr. K. Ramesh Kumar, N. Monica, (2014), "Survey on mining clusters using new k-mean algorithm from structured and unstructured data T" pp. 60-65.

[10] 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.

[11] Pokrajac, Dragoljub, Aleksandar Lazarevic, and Longin Jan Latecki, 2007. "Incremental local outlier detection for data streams." Computational Intelligence and Data Mining, 2007. CIDM 2007. IEEE Symposium on. IEEE.

[12] Sreevidya S. S. *et al*., (2014), "A Survey on Outlier Detection Methods" International Journal of Computer Science and Information Technologies, (IJCSIT) Vol. 5 (6), 13. pp. 8153-8156.