International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-7, Issue-6, March 2019

Outlier Detection in Imbalanced Data Classification

M. Kamaladevi, K. R. Sekar, V. Venkataraman, K. Kannan

Abstract: ***In Binary classification , the distribution of classes present in a data is not uniform such that the number of instances of a class(es) significantly out numbers the instances of another class(es) leads to class imbalance. Classification algorithm biased toward the majority class. Performance accuracy are not based on minority class instance. This lead to degrade the classifier .To improve performance characteristics of minority data instance such as borderline rare and outlier has to analyzed. An outlier or an anomaly is a point that deviates from the normal behavior exhibited by the other points in a data. Detection of outlier in class instances is still open Research. Problem. In this article, two density-based outlier detection methods are compared. The two methods in discussion are the KNN method and the Local Outlier Factor (LOF). The KNN algorithm, which is a classification algorithm, is a global density-based method, while the LOF is a local density-based method. These two methods are applied on the imbalanced data set Breast Cancer-W Dataset, consisting of 569 instances and 33 variables, taken from the UCI (University of California, Irvine) Machine Learning repository. The accuracy of both the algorithms (based on the percentage of observations correctly identified) is found out and their performances are analyzed. It has been found out that LOF method provided a better view of outlier data compared to KNN method.***

Key Words: Outlier, LOF, KNN, distance-based, density-based

I. INTRODUCTION

Existing classifier work with an assumption that class in the data are balanced. Their training performance is estimated using predictive accuracy or 0-1 loss function, both are assume the data set distribution is uniform. But this is not valid when the class distribution is not uniform or imbalanced. Data of Majority class instance is more than minority class instance. Classification accuracy is also depend on minority class instance and considering those instance have an impact on some real time scenario such medical diagnosis, fraud detection. Analyzing the Characteristics of minority class instance and their influence on classification performance is still a open problem in imbalance classification. Minority class considered to fall with this four types: safe , borderline, rare and outlier. The cause of outliers could be because of error in dataset, measurement error or correct but exceptional data. The detection of outliers is useful in applications like network intrusion, credit card fraud etc..

Revised Manuscript Received on 30 March 2019.
* Correspondence Author
M. Kamaladevi, School of Computing ,SASTRA Deemed University, India
K.R.Sekar, School of Computing ,SASTRA Deemed University, India
V.Venkataraman, School of Humanities and Science, SASTRA Deemed University, India
K.Kannan, School of Humanities and Science, SASTRA Deemed University, India

Various methodologies have been discussed in the past to detect outliers (Statistical, distance-based, density-based, etc.) Outlier data need to be isolated first from the set of observations, and the definition what constitutes data outlier need to be context-based on the nature of the data. Once the definition is made, then the next question is how to establish the data outlier present. This may call for application of certain methods that can be used to isolate the data outlier.
It is also important to identify outlier either with independent data sets, or determine outlier based on combination of data sets. Individually, the data item may be within the range, but when used in conjunction with other data variables, a particular value may be considered as an outlier. There are multiple approaches to identify outliers as follows:
The first set of approaches are based on data model and include the following:

- Statistical Methods
- Depth based approaches
- Deviation based approaches

The second set of approaches involve proximity and deal with:

- distance based approaches
- density based approaches

Compared to the statistical approach, distance based approach measures the distance between the point to its neighbors. There are multiple algorithms in measuring the distance based outliers: index-based, nested loop and cell-based. Density based outlier detection algorithms are based on the proximity from one point to its neighbors in the dataset. This approach is also proximity-based and is considered a better model, compared to distance based outlier detection. The basic theme of this model is the use of the LOF (Local Outlier Factor) concept. LOF compares the density around a point with the density around its local neighbors.

II. .RELATED WORK

Outlier Detection for temporal data has been been a analyzed using density based , distance based and network approaches[1]. Probabilistic approach using SVM on biomedical data, generating weak outliers at alert rates of 25% and stronger outliers at 66%. Outlier detection method by first computing likelihood values using k-means clustering and LOF algorithms, thus generating a dataset. Then the general likelihood values are used with abnormal examples using the SVDD algorithm to build an accurate classifier for outlier detection.[2] Outlier detection techniques i.e based on supervised and unsupervised techniques and their performance on dynamic and static graphs, and discuss their effectiveness, scalability, robustness on graphical data.[3].

Data preprocessing in minority class instance has play a significant role in classification accuracy. Identifying outlier data in minority class instance is the important criteria[6]. Imbalanced Data will get best classification accuracy using SVM and KNN algorithm[5]. Four types of anomaly detection: Clustering, Information theory, Classification and Statistical based methods. are analyzed for outlier detection [4].Minority class instance example are preprocessed using synthetic minority oversampling[8].Outlier detection and spatial method discussed in it, first breaks the system calls into multiple n-grams, with varied length, and applying OC-SVM with a Gaussian Kernel [9][10].

III. APPLIED METHODLOGY

As regards this aritcle, two density based outlier detection algorithms have been applied for imbalanced data set

- KNN algorithm (Manhattan metric) in conjunction with statistical approaches
- Local Outlier Factor (LOF) approach

The KNN algorithm, which is a classification algorithm, works by choosing k number of neighbors and computing the distance of the point to its k neighbors. The point is labeled based on the labels of majority of the points close to it. The higher distance of a point to its KNN, the lower the local density and hence the point is most likely an outlier. In this work, the data is first classified using the K-NN approach. Then the outliers in the data are detected by a statistical outlier detection approach (by considering values outside the 1.5 sigma limit to be outliers). The Local outlier Factor algorithm is based on the concept of the reachability distance, which is the maximum of the distance of the point to its k-th nearest neighbor and Euclidean distance of point to another point. These distances are then averaged and the inverse is found out. This resulting quantity is called the Local outlier Factor Score. This is found out for all the other points. Local outlier Factor scores close to 1 is an outlier, while scores higher than 1 mean that the point is an outlier.

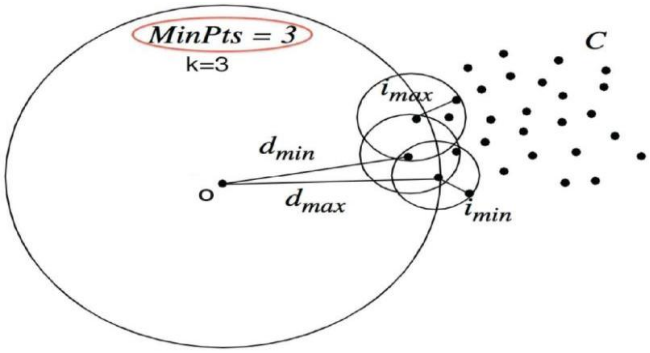


Fig 1.1: Diagrammatic description of the LOF algorithm with k=3

The mathematical working of the LOF algorithm is specified below:

Reachibility distance(A,B)=max{k-distance(B), distance(A,B)}

The local reachability density of an object A is defined by

$Ird(A):=1/ (\sum_{B\in} N_{k\,(A)}$-reachbility-distance(A,B) $/|N_k(A)|)$

The local reachability densities are then compared with those of the neighbors using

$LOF_k(A)= ((\sum_{B\in} N_{k\,(A)}$-Ird(B) / Ird(A)) $/|N_k(A)|)$

$= ((\sum_{B\in} N_{k\,(A)}$-Ird(B)) $/|N_k(A)|)/Ird(A)$

At first the minority class (Malignancy) data is split into the training and the test dataset. Then the variables that correlate the most with the class variable are chosen for analysis. The training data is then classified using the KNN algorithm. The Outlier is detected by number of observation fall within 1.5 sigma limit. And accuracy is found by dividing the number of outlier in the test data by the total number of observations, which is expressed as a percentage. The Local outlier Factor score is calculated for all the points in the test data. The number of points whose LOF score fall within the threshold is identified as outlier. Then the accuracy of the LOF algorithm is found out by dividing the number of such points by the total number of observations and expressing it as a percentage. Then the output as to which algorithm is better than the other is displayed based on the comparison of their accuracies.

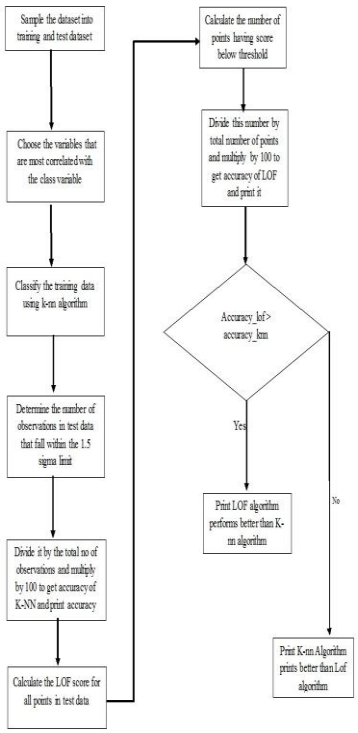
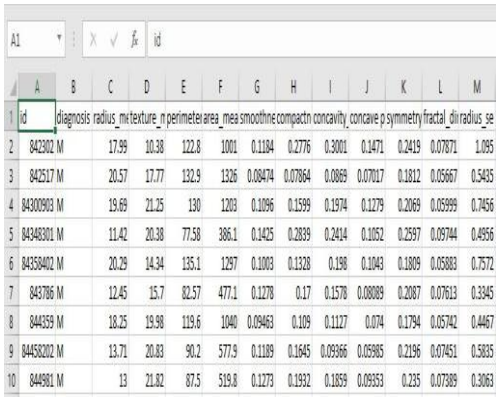


Fig 1.2 Flowchart for Outlier detection

The most important variables from the dataset are first chosen by correlation analysis. The extracted features (or variables) are then used in building the training and test data. 70% of the data have been treated as training set, and the balance as test data. Then, the KNN classification algorithm and the LOF algorithm, the two outlier detection algorithms, are used on the training and test data after which the output is displayed to the user.

The dataset titled "Breast Cancer Wisconsin Dataset" was taken from the UCI Machine Learning Repository considered as binary imbalanced classification of imabalanced Ratio 2.33. Malignance is considered as minority class and benign as majority class. This dataset consists of 569 observations and 33 variables that consist of various dimensions of the cancer cells like area, radius, fractal dimension, symmetry, perimeter etc. The class variable denotes the severity of cancer (M-Malignant, B-benign). The snapshot of the sample data set is provided below:



id	diagnosis	radius_me	texture_m	perimeter	area_mea	smoothne	compactn	concavity	concave p	symmetry	fractal_di	radius_se
842302	M	17.99	10.38	122.8	1001	0.1184	0.2776	0.3001	0.1471	0.2419	0.07871	1.095
842517	M	20.57	17.77	132.9	1326	0.08474	0.07864	0.0869	0.07017	0.1812	0.05667	0.5435
84300903	M	19.69	21.25	130	1203	0.1096	0.1599	0.1974	0.1279	0.2069	0.05999	0.7456
84348301	M	11.42	20.38	77.58	386.1	0.1425	0.2839	0.2414	0.1052	0.2597	0.09744	0.4956
84358402	M	20.29	14.34	135.1	1297	0.1003	0.1328	0.198	0.1043	0.1809	0.05883	0.7572
843786	M	12.45	15.7	82.57	477.1	0.1278	0.17	0.1578	0.08089	0.2087	0.07613	0.3345
844359	M	18.25	19.98	119.6	1040	0.09463	0.109	0.1127	0.074	0.1794	0.05742	0.4467
84458202	M	13.71	20.83	90.2	577.9	0.1189	0.1645	0.09366	0.05985	0.2196	0.07451	0.5835
844981	M	13	21.82	87.5	519.8	0.1273	0.1932	0.1859	0.09353	0.235	0.07389	0.3063

Fig 1.3 Sample Data Set

IV. RESULT AND DISCUSSION

The histogram plot of the KNN algorithm (with Manhattan metric)(Fig 1.4) along with the density plot of the LOF algorithm(fig 1.5) and the most correlated variables (area,radius,symmetry,perimeter) are displayed. The accuracy of both the algorithms and the result are displayed in the figure given below

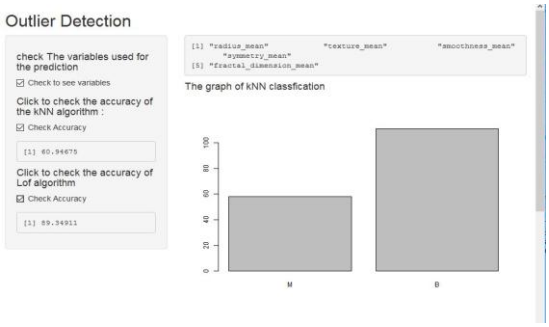


Fig 1.4 Histogram Plot of KNN Algorithm

The methods could show their relative merits in detecting the outlier data. Accuracy of both methods (expressed in % of observations identified correctly) is presented below.

Table1.1: Accuracy Performance of the KNN and

Outlier Detection Method	Accuracy(%)
K NN Algorithm (ManhattanMetric)	60.94
LOF algorithm	89.34

Local outlier Factor(LOF) Algorithms

Local outlier Factor giving more accuracy for finding outlier than KNN in imbalanced dataset. For minority class instances Local outlier Factor algorithm to be better for finding outlier.

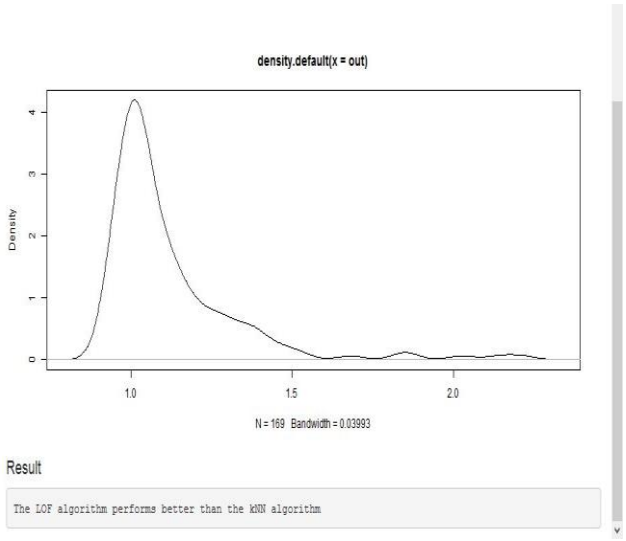


Fig1.5 Density plot of LOF Algorithm

V. CONCLUSION

In Imbalanced Data Classification ,outlier present in the instances of the given data set is identified using KNN algorithm and Local outlier Factor(LOF) algorithm. The accuracies of both algorithms as seen from Table 1.1 shows that Local outlier Factor (LOF) has a higher accuracy than K-NN algorithm. The percentage of observations identified by the Local outlier Factor(LOF) algorithm is higher than that of the KNN algorithm (with Manhattan metric).

This is because, the working of the Local outlier Factor(LOF) algorithm is based on the concept of local density, wherein a point being anomalous with respect to a set of points, is not ruled out right away as an outlier. Hence, this algorithm works well for varying densities. Whereas in the KNN algorithms, global density of the points is used. So it bring down the accuracy.

Though number of methods have evolved to detect the outlier, the methods used may suffer due to dimensions (curse of dimensionality). In future ,outlier detection can be done using newer methods that can handle volume of data, and number of dimensions. The methods include:

- Angle based outlier degree (ABOD)
- Grid-based subspace outlier detection

ACKNOWLEDGEMENT

The authors of this paper wish to thank Prof R. Sethuraman, Vice Chancellor and Dr. S. Swaminathan, Dean(Sponsored Research) of our University for giving permission to utilize the DST-FIST sponsored Discrete Mathematics Laboratory at Srinivasa Ramanujan Centre, Kumbakonam to do this research work

REFERENCES

1. Gupta, M., Gao, J., Aggarwal, C., & Han, J. (2014). Outlier detection for temporal data. *Synthesis Lectures on Data Mining and Knowledge Discovery*, *5*(1), 1-129.
2. Liu, B., Xiao, Y., Philip, S. Y., Hao, Z., & Cao, L. (2014). An efficient approach for outlier detection with imperfect data labels. *IEEE transactions on knowledge and data engineering*, *26*(7), 1602-1616.
3. 3.Akoglu, L., Tong, H., & Koutra, D. (2015). Graph based anomaly detection and description: a survey. *Data mining and knowledge discovery*, *29*(3), 626-688.
4. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, *60*, 19-31.
5. Wah, Y. B., Rahman, H. A. A., He, H., & Bulgiba, A. (2016, June). Handling imbalanced dataset using SVM and k-NN approach. In *AIP Conference Proceedings* (Vol. 1750, No. 1, p. 020023). AIP Publishing.
6. 6..Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221-232.
7. Napierala, K., & Stefanowski, J. (2016). Types of minority class examples and their influence on learning classifiers from imbalanced data. *Journal of Intelligent Information Systems*, *46*(3), 563-597.
8. 8.. Skryjomski, P., & Krawczyk, B. (2017, October). Influence of minority class instance types on SMOTE imbalanced data oversampling. In *First International Workshop on Learning with Imbalanced Domains: Theory and Applications* (pp. 7-21).
9. 9 .Bosman, H. H., Iacca, G., Tejada, A., Wörtche, H. J., & Liotta, A. (2017). Spatial anomaly detection in sensor networks using neighborhood information. *Information Fusion*, *33*, 41-56.
10. Khreich, W., Khosravifar, B., Hamou-Lhadj, A., & Talhi, C. (2017). An anomaly detection system based on variable N-gram features and one-class SVM. *Information and Software Technology*, *91*, 186-197.