Comparison of Outlier Detection Methods in Diabetes Data

V. Mahalakshmi
Assistant Professor,
Department of Computer Science and Engineering,
Annamalai University,
Annamalai Nagar – 608002,
Tamilnadu, India.

M. Govindarajan
Assistant Professor,
Department of Computer Science and Engineering
Annamalai University,
Annamalai Nagar – 608002,
Tamilnadu, India.

ABSTRACT

Outlier is defined as an observation that deviates extensively from other observations. The identification of outliers can lead to the discovery of useful and meaningful knowledge. Outlier detection has been widely studied in the past decades. Most refined methods in data mining address this issue to some extent, but not fully, and can be improved by addressing the problem more directly. The detection of outliers can lead to the invention of unpredicted facts in areas such as credit card fraud detection, calling card fraud detection, discovering criminal behaviors, discovering network intrusions, etc. This paper mainly discusses and compares approach of different outlier detection from data mining perspective, which can be grouped into distance-based approach and density-based approach.

Keywords

Outlier detection, Distance-based approach, Density-based approach

1. INTRODUCTION

Data mining is a process of extracting valid, previously unidentified, and ultimately understandable information from huge datasets and using it for organizational decision-making. However, there are a lot of problems that exist in mining data in large datasets such as data redundancy, the value of attributes being not specific, data is not complete and presence of outliers. There is large amount of data that is available in information industry. This data can be utilized only when it is converted into useful information. It is necessary to analyse this large amount of data and extract useful information from it.

Outlier is defined as an observation that deviates too much from other observations that it arouses suspicions that it was generated by a different mechanism from other interpretations [1]. The detection of outliers can lead to the discovery of useful knowledge and has a number of practical applications in areas such as public safety, transportation, public health and location based services. This paper mainly discusses about outlier detection approaches from data mining perspective. The inherent idea is to analyze and compare achieving mechanism of those approaches to determine which approach is better based on noisy datasets.

In Section 2 related work is discussed. Description about the dataset is given in section 3. Section 4 presents the various outlier detection methods employed. The performance of outlier detection methods employed is discussed in Section 5. Section 6 concludes the paper.

2. RELATED WORK

Distance-based outlier methods have time and space complexities [2]. I-CLARANS algorithm proposed by [1] considers three existing partition based clustering algorithms called PAM, CLARA and CLARANS and also combines them with distance-based method for outlier detection. This reduces computation time considerably. The I-CLARANS identifies outliers more successfully than existing algorithms [3].

Hybrid approach proposed in [1] contains cluster-based approach and distance based approach. Hybrid approach reduces the time and space complexity. However, there is no mention of any impact on cost and performance by using hybrid technique [1]. Dutta et al. [4] proposed algorithms for the distributed computation of principal components and top-k outlier detection. In their approach, outliers are defined as objects that distinguish themselves from the correlation structure of the data.

The applications of data streams generated in transactions, ATM data, credit card operations and popular web site logs led to the study of outlier detection in data stream [5]. The approach proposed by Angiulli et al [6] considered a weighted sum of the distances from the k nearest neighbours to each data point, and classifies as outliers those points which have the largest weighted sums. The basic idea of density-based approaches is that the density around an outlier remarkably varies from that around its neighbors [7].

Any appropriate distance measure can be used such as Euclidean distance, Mahalanobis distance, or some other measure of dissimilarity. Usually, the choice of distance measure depends on the type of the variables. Chawla and Gionis [8] presented a technique which simultaneously clusters and discovers outliers in data. This approach is the generalization of K-means approach and hence it is NP-Hard. It is an iterative approach and it converges to local optima.

3. DATASETS DESCRIPTION

The dataset chosen for this work is PIMA Indian Diabetes dataset because it has been widely studied and is considered a difficult set. There are 268 (34.9%) cases in class ‘1’ and 500 (65.1%) cases in class ‘0’. Table 1 represents the description of dataset used.

Table 1 Description of PIMA Diabetes dataset

Properties	Value
Number of Samples	768
Number of attributes	8

Number of classes	2
Type of attributes	Numeric
Type of Class attributes	Binomial

3.1. PIMA Noise Datasets

The standard classification task consists of making generalizations from a set of training examples.

Table 2 Types of PIMA noise datasets

Datasets used	% of noise added	Description
Pima Noise Dataset 0 (PND0)	0%	Original dataset
Pima Noise Dataset 1 (PND1)	5%	5% attribute noise added
Pima Noise Dataset 2 (PND2)	10%	10% attribute noise added
Pima Noise Dataset 3 (PND3)	15%	15% attribute noise added
Pima Noise Dataset 4 (PND4)	20%	20% attribute noise added

The attribute noise is introduced in the data set. Five different types of datasets namely the original dataset, and datasets with 5%, 10%, 15% and 20% attribute noise are used as shown in Table 2.

4. OUTLIER DETECTION METHODS

Outlier detection is a preprocessing stage for knowledge extraction. When data is found to be free of outliers on analysis there will be a reduced level of ambiguity and fuzziness. Experiments were performed on several data sets summarized in Table 2. In all the experiments, it is assumed that the information about the normal behavior (class) is provided in the data set. In this work, two different techniques are used for detecting outliers. The performance of these techniques were evaluated. Density based, and distance-based techniques are employed in the above mentioned data sets. These approaches tend to assign each object with a value that measures the degree to which it is an outlier.

4.1. Distance based Outlier Detection

Various distance-based methods are available, of which k-NN distance-based outlier detection technique is used in this work. Outlier detection is based on the distance of an object to its k nearest neighbor. In this implementation, the k nearest neighbors does not exclude the point that is currently evaluated. For each object in the dataset, the number of objects that lie within a distance is recorded. Based on the distance criteria, the number of objects that have least number of neighbors within a specified radius are considered to be noisy. The objects will be sorted in ascending order based on the number of neighbors it has. The first *n* outliers are considered as noise and are removed from the dataset. The parameters used for k- NN distance-based outlier detection technique is distance function and value of ‘k’. Euclidean distance function is used.

Fig 1 show that the outlier values are distributed between 10 and 300. In Fig 1, for PND1, the maximum number of instances are distributed around outlier score of around 20-40 for both positive and negative class instances. The peak occurs at outlier score value of 23.

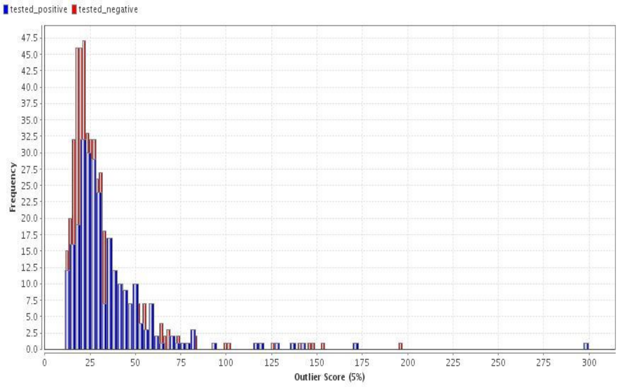

Fig 1 Distance based outlier score distribution of PND1

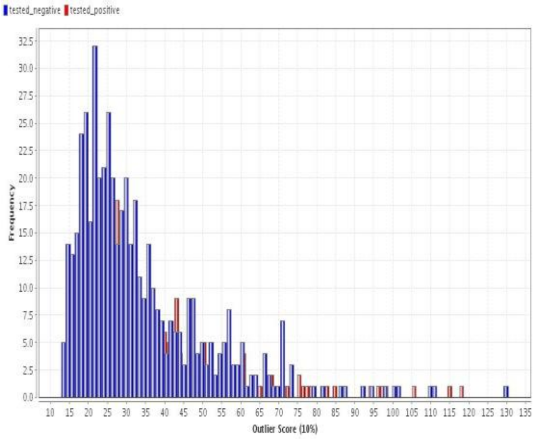

Fig 2 Distance based outlier score distribution of PND2

From Fig 2, it can be inferred that the outlier values are distributed between 13 and 130. In Fig 2, for PND2, the maximum number of instances are distributed with outlier score of around 15-55 for both positive and negative class instances. Among the outlier scores calculated, the top 10 outliers are to be identified from outlier score greater than 74.

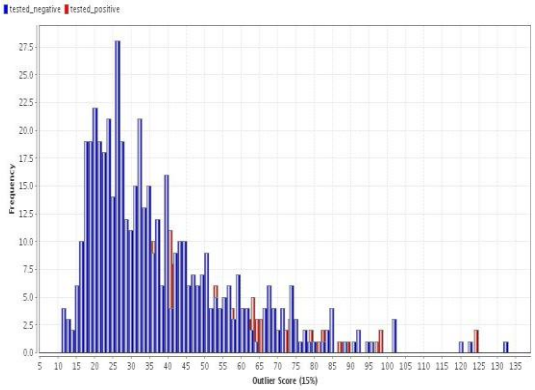

Fig 3 Distance based outlier score distribution of PND3

For PND3 the distance based outlier detection technique generates outlier scores ranging between 11 and 132. The resultant histogram for PND3 is shown in Fig 3. Among the outlier scores calculated, it can be seen that very few instances have highest outlier score of 132.

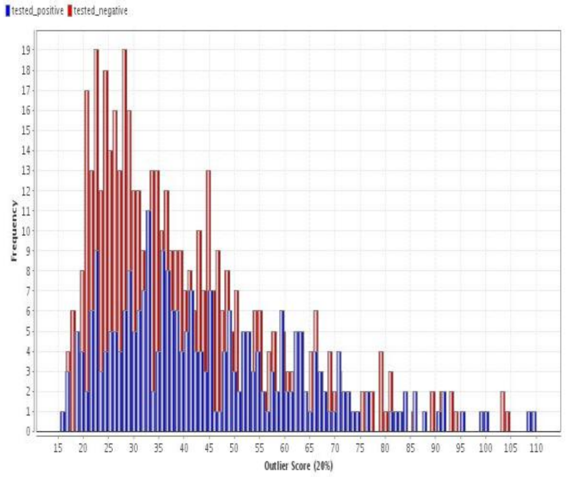

Fig 4 Distance based outlier score distribution of PND4

For PND4 the distance based outlier detection technique generates outlier scores ranging between 15 and 110. The resultant histogram for PND4 is shown in Fig 4. In Fig 4, for PND4, the maximum number of instances are distributed evenly for both positive and negative class instances. Thus, both class instances have been identified as outliers. The peak occurs at outlier score value of 28.

4.2. Density based Outlier Detection

Density based outlier detection uses density distribution of data points within the data set. Outliers in this technique are measured by using a local outlier factor (LOF). LOF is the ratio of local density of an object and the local density of its nearest neighbor. Outliers here are data objects with higher LOF. The algorithm to compute density-based local outlier factor in a database is based on minPts. The original LOF parameter was called "minPts", but for consistency within ELKI this parameter is named as "k". The ‘k’ value represents the number of nearest neighbors whose distance is used to estimate the density. The distance function used in this approach is Euclidean function. For the datasets, PND1, PND2, PND3, and PND4 used, the calculated outlier values are sketched in Fig 5 – Fig 8 in the form of a histogram.

Fig 5 shows the histogram as a distribution plot for PND1 dataset. Density based outlier detection rather presents a different distribution. Fig 5 shows that outlier values are distributed between 1 and 5.6. The peak occurs at outlier score value of 1.1. Among the outlier scores calculated, it can be seen that only one instance has a highest outlier score of 5.6.

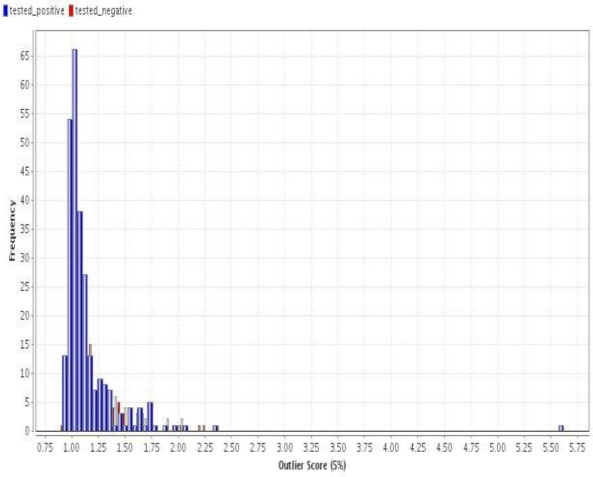

Fig 5 Density based outlier score distribution of PND1

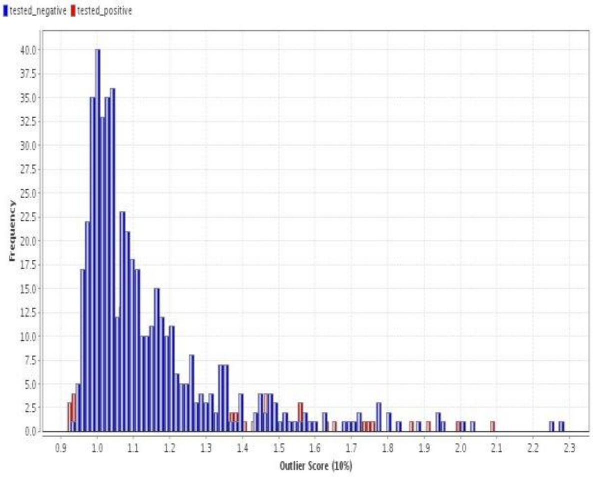

Fig 6 Density based outlier score distribution of PND2

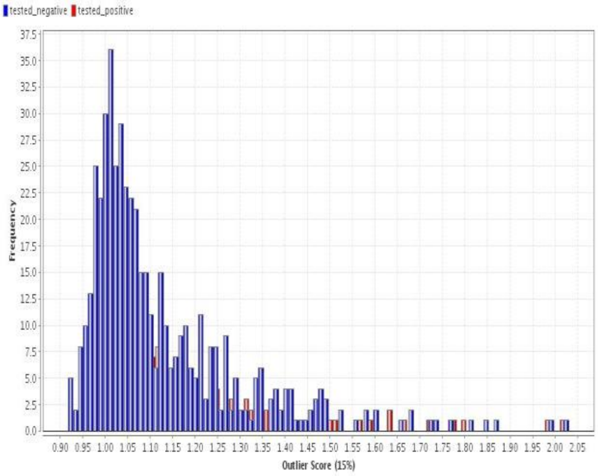

Fig 7 Density based outlier score distribution of PND3

From Fig 6, it can be inferred that the outlier values are distributed between 0.9 and 2.3. In Fig 6, for PND1, the maximum number of instances are distributed around outlier score of around 1-1.8 for both positive and negative class instances. Among the outlier scores calculated, the top 10 outliers are to be identified from outlier score greater than 2.1. The number of instances identified with higher outlier scores using density based outlier detection is much higher for PND3 than PND2 (Fig 6). The instances with lower outlier score in the range of 0.9-1.7 are normal instances. Among the outlier scores calculated, the top 10 outliers whose outlier score greater than 1.86 are to be identified as shown in fig 7.

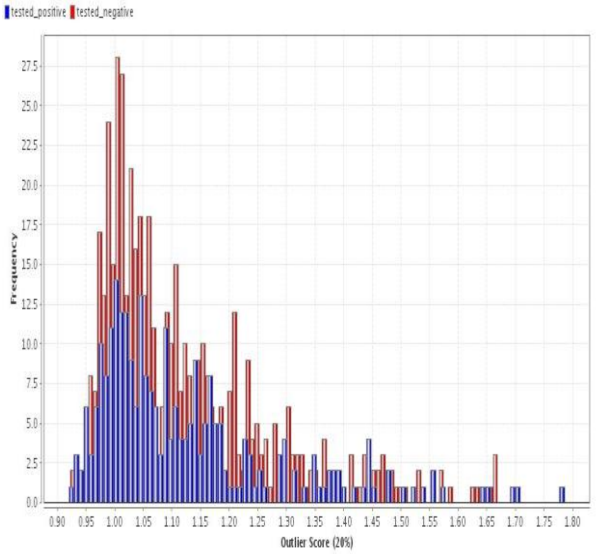

Fig 8 Density based outlier score distribution of PND4

For PND4, the density based outlier detection technique generates outlier scores ranging between 0.93 and 1.78. The resultant histogram for PND4 is shown in fig 8. Among the outlier scores calculated, the top 10 outliers whose outlier score greater than 1.66 are to be identified. For PND4 dataset, the number of instances with higher outlier scores is more when compared to other datasets.

5. PERFORMANCE OF OUTLIER DETECTION METHODS

Machine learning has various methods for evaluating the performance of learning algorithms. After employing the distance based, and density based outlier detection methods on PND1, PND2, PND3 and PND4 datasets, the top ten outliers are removed. After removal, the number of instances in the datasets is reduced to 758 from 768. Then the performance is measured by making use of a classifier. To compare the performance of various outlier detection techniques, two different classifiers are used. The classifiers are Support Vector Machine (SVM), and Naïve Bayes (NB). These classifiers are employed much in the literature of outlier detection [9].

Evaluation measures analyse different characteristics of machine learning algorithms. Accuracy measure describes the extent to which the set of tuples are classified correctly. Classification accuracy is one of the most popular metrics used in the evaluation of classifiers. The performance measure of the classifiers is also evaluated by employing original clean PIMA dataset.

Table 3 Performance measure for PND0

Performance measure	**Performance (%)**	
	NB	**SVM**
Accuracy	68.3	77.4

The classification results obtained for the classifiers Support vector machine and Naïve Bayes are shown in Table 3 for the actual datasets. The results of SVM (Table 3) show that the accuracy is comparatively larger than that of Naive Bayes classifier.

5.1. Accuracy of classifiers

Table 4 shows the accuracy values of classifiers for outlier detection methods for all noisy datasets used. From table 4, it is observed that the performance of classifiers varies depending on the outlier detection methods employed. Among the different outlier detection methods, distance based outlier detection performs better for PND1 dataset. The hybrid combination of SVM and distance based OD method has greater accuracy of 75.2% among all other hybrid combination of classifiers and outlier detection methods. In general, the results show that SVM performs better for a 5% PIMA noise dataset (PND1).

With SVM as classification method, the density based outlier detection method has greater accuracy of 75.2% for a 10% PIMA noise dataset (PND2). It is also observed that, all classifiers employed has its classification accuracy improved when compared to the accuracy obtained before removing outliers for a 10% PIMA noise dataset (PND2). PND3 Thus, it is observed from results in table 4 that the compound combination of SVM and density based outlier detection method gives better accuracy for 15% PIMA noise dataset (PND3). Thus, from the results in Table 4 it is inferred that SVM performs better for all datasets invariant of the type of outlier detection method employed.

6. CONCLUSION

Outlier detection is a broad field, which has been studied in the context of a large number of application domains. Outlier detection algorithms aim to identify valuable and disturbing observations in large collections of data. In this work, the performance of outlier detection approaches for PIMA diabetes dataset is analysed. The efficiency of outlier detection is evaluated by introducing noise to the actual dataset at various levels. The performance is evaluated by implementing two classifiers. For all the four datasets, SVM classifier is found to generate better performance accuracy. For PND1 dataset, distance based outlier detection method gives more accuracy than density based outlier detection method. But for the remaining three datasets, PND2, PND3, and PND4, density based outlier detection method generates more accuracy. An overall study of accuracy measure for all the four datasets shows that density based outlier detection method performs better than the other outlier detection methods employed.

Table 4 Accuracy of classification methods for all datasets

Method/Classifier	**PND1**		**PND2**		**PND3**		**PND4**	
	NB (%)	**SVM (%)**	**NB (%)**	**SVM (%)**	**NB (%)**	**SVM (%)**	**NB (%)**	**SVM (%)**
Before outlier detection	68.3	74.8	70.9	74.1	66.5	72.3	73.1	72
Distance based	70.9	*75.2*	70.9	74	70.1	71.9	69.6	71.7
Density based	69.2	74.9	71.8	*75.2*	73.1	*75.7*	73.7	*76*

7. REFERENCES

[1] Surekha V Peshatwar & Snehlata Dongre, "Outlier Detection Over Data Stream Using Cluster Based Approach And Distance Based Approach", *International Conference on Electrical Engineering and Computer Science (ICEECS-2012),* Trivandrum, May 12th, 2012.

[2] Pranjali Kasture, Jayant Gadge, "Cluster based Outlier Detection", *International Journal of Computer Applications (0975 – 8887)* Vol.58(10), November 2012.

[3] Garima Singh, Vijay Kumar, "An Efficient Clustering and Distance Based Approach for Outlier Detection", *International Journal of Computer Trends and Technology (IJCTT),* Vol.4(7), July 2013.

[4] H. Dutta, C. Giannella, K.D. Borne, and H. Kargupta, "Distributed Top-K Outlier Detection from Astronomy Catalogs Using the DEMAC System," *Proc. SIAM International Conference in Data Mining (SDM),* 2007.

[5] C. Aggarwal, J. Han, J. Wang, P.S. Yu, "A framework for projected clustering of high dimensional data streams", in Proceedings of the 30th VLDB Conference, Toronto, Canada, 2004, pp. 852-863.

[6] Angiulli, F., and Pizzuti, C., "Fast outlier detection in high dimensional spaces", *Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery, PKDD 002*, Springer-Verlag, London, UK, 2002, pp. 15–26.

[7] Chandola, V., Banerjee, A., and Kumar, V., "Anomaly detection: a survey", *ACM Comput. Surv. (CSUR),* Vol. 41(3), 2009, pp. 1–58.

[8] Chawla, S., and Gionis, A., "k-means–: A unified approach to clustering and outlier detection", *SDM, SIAM*, 2013, pp. 189–197.

[9] Acuna, E., and Rodriguez, C. A., "A meta analysis study of outlier detection methods in classification", *Technical paper,* Department of Mathematics, University of Puerto Rico at Mayaguez, 2004, pp. 1-25.