A Performance Analysis of the Innovative Methods Employed for Outlier Detection using Data Mining Algorithms with Three Different Applications

R Delshi Howsalya Devi and Dr. M Indra Devi

1Assistant Professor, K.L.N College of Engineering, Madurai, India.
2Professor/HOD, Kamaraj College of Engineering, Madurai, India.

Received 25 April 2016; Accepted 28 June 2016; Available 7 July 2016

ABSTRACT
Data Mining simply refers to the mining of very interesting patterns of the data from the massive data sets. Outlier detection is one of the important characteristics of data mining. It is a task that finds objects that are considerably dissimilar, incomparable or inconsistent with respect to the remaining data. Outlier detection has wide applications which include data analysis, network intrusion detection, financial fraud detection, and clinical diagnosis of diseases. This paper proposes three outlier detection models such as OFWDT (Outlier Finding with Decision Tree), OFWNB (Outlier Finding with Naive Bayes) and OFWQR (Outlier Finding With Quartile Range) with three different applications. OFWDT model has three steps of a process. In the first step, groups the data in to number of clusters using Farthest First clustering algorithm. Due to minimize the size of dataset, the computation time reduced greatly. In the second step, outliers are detected from wisconsin breast cancer dataset using ODA (Outlier Detection Algorithm). In third step, identifies whether the cancer is malignant or benign from the pre-processed data set using J48 classification algorithm. OFWNB model consists of two phases. In first phase, the Farthest First (FF) clustering algorithm is executed. Small clusters are determined and considered as outlier clusters. In second phase, the rest of outliers (if any) are detected in the remaining clusters based on computing the absolute distances between the Centroids and Points [ADCP]. In each cluster a threshold value is also calculated. If the ADCP value is less than the threshold value the object is retained in the cluster, but, if the ADCP value is greater than the threshold value the object is detected as outlier. Finally, the preprocessed data is then classified either good or bad using Naive Bayes classifier. OFWQR model based on a hybrid method that uses IQR filter for preprocessing the original data set by removing outliers and j48 decision tree classifier is applied for diagnose the heart data into healthy and a patient who is focus to possible heart disease. The aim of this paper is to study the performance comparison of three outlier detection methods with three different applications such as breast cancer diagnosis, Credit card fraud detection and Heart disease diagnosis. We used the Wisconsin Breast cancer data set for diagnose the breast cancer. Credit card german data set is used for fraud detection and Cleveland heart disease data set is used for heart disease diagnosis. The experimental results proves that the outlier finding with decision tree (OFWDT) model obtained 99.9%,99.8%,99.9% classification accuracy for diagnosis of breast cancer, credit card fraud detection and heart disease diagnosis. OFWDT model gives better performance compared with three applications.

KEYWORDS: Outlier Detection, Credit card, Heart disease, Breast Cancer, IQR filter,NB,J48

INTRODUCTION

Nowadays million of databases have been used in business management, Govt., scientific engineering & in many other application & keeps growing rapidly in present day scenario. The explosive growth in data & database has generated an urgent need to develop new technique to remove outliers for effective data mining.
Outliers can significantly affect data mining performance. It is an important issue in knowledge discovery and data mining and has attracted increasing interests in recent years. Outlier detection has important applications in many fields in which the data can contain high dimensions. The task of outlier detection is to find the small groups of data objects that are exceptional to the inherent behavior of the rest of the data. Detection of such outliers is fundamental to a variety of database and analytic tasks such as fraud detection and customer migration. There are several approaches [10] of outlier detection employed in many study areas amongst which distance based and density based outlier detection techniques have gathered most attention of researchers. In this paper investigate a novel approaches for outlier detection in three application such as breast cancer, heart disease diagnosis and credit card fraud detection. Breast cancer is the most common malignancy type among the female population, particularly in industrialized countries. Medical treatment of this disease is based on diagnosis and prognostic factors. Several computer aided systems and methods have been presented in order to obtain high classification accuracy. Motivated by this need, in this paper implemented a three machine learning experiments on the WBCD dataset by using different approaches in order to achieve better prognosis predictions. Fraud is one of the major causes of great financial losses, not only for merchants, but individual clients are also affected. Many data mining techniques have evolved in detecting various credit card fraudulent transactions. Out of the available data mining techniques outlier detection with machine learning classification algorithms is used to find these mischief activities with better classification accuracy. Various classification and regression processes have been used to identify heart disease. In particular, focus has been made on the University of California Irvine (UCI) heart disease dataset (also known as the Cleveland dataset and different outlier detection model with classification algorithms. have been used. But, existing investigations are, to the best of the author’s knowledge, yet to show a comparative research that considers three different outlier model achieves better classification accuracy for predict heart disease.

The rest of the paper is organized as follows. Section II provides discussion on the previous works related to the topic. Section III describes briefly outlier detection models: OFWDT, OFWNB and OFWQR. Section IV describes the Proposed models for outlier detection. Section V describes experimental study and results. Conclusion is given in Section V.

**Related Work:**

There is a lot of literature on the outlier detection methods with diagnosis of breast cancer, heart disease and credit card fraud detection problem which describes a variety of approaches like clustering, density based outlier detection, distance based outlier detection. Inan, Uzer, and Yilmaz (2013) has achieved 98.29% classification accuracy with an integrated model of association rule mining based feature selection, the principal component analysis and a neural network classifier[6]. Uzer, Inan, and Yilmaz (2013) have proposed the integrated approach of principal component analysis with a sequential forward selection and sequential backward selection based feature selection method and has obtained 98.57% classification accuracy. Seera and Lim (2014) has obtained a classification accuracy of 98.84% for breast cancer diagnosis with a hybrid intelligent classification model for medical data. The model contains the Fuzzy Min–Max neural network, the classification and regression tree and the Random Forest algorithm. The Fuzzy Min–Max neural network is dependable for incremental learning, the classification and regression tree is responsible for attractive understandability and the Random Forest algorithm is utilized to enhance the predictive performance[7]. Li, Peng, and Liu (2013) have discussed the quasi formal kernel common locality discriminant analysis for dimensionality reduction and have reported a classification accuracy of 97.26%. Zheng, Yoon, and Lam (2014) have achieved a classification accuracy of 97.38% with K-means algorithm and support vector machine based model[8]. Ozsen and Ceylan (2014) presented the performance of artificial immune system as a data reduction algorithm. In order to estimate the data reduction performance of artificial immune system, it is evaluated to the fuzzy c-means clustering algorithm. Both data reduction algorithms are combined with the artificial neural network classifier to get the classification results. They have achieved 97.80% classification accuracy for artificial immune system and artificial neural network combination, whereas a c fuzzy c-means clustering and artificial neural network combination they have obtained classification accuracy of 90.04%[9].

Chen (2014) has proposed a combined model for breast cancer diagnoses that can work in the absence of labeled training data. Hence, this work explains the feature selection methods in unsupervised learning models. The model combines clustering and feature selection. Their work indicates that selecting a subset of relevant features instead of using all the features in the original data set can improve the interpretability of clustering results[10]. Nguyen, Khosravi, Creighton, and Nahavandi (2015a) have obtained a classification accuracy of 97.88% with a medical classification model which integrates wavelet transformation and interval type-2 fuzzy logic system. These mechanism are combined together in order to increase the dimensionality and uncertainty properly. Interval type-2 fuzzy logic system consists of fuzzy c-means clustering based unsupervised learning and genetic algorithm based parameter tuning. These mechanisms have high computational costs and wavelet transform functions for reducing these computational costs[11]. Bhardwaj and Tiwari (2015) have achieved a classification accuracy of 99.26% for 10-fold cross validation scheme. They have proposed a genetically
optimized neural network algorithm for medical diagnosis. This paper extends crossover and mutation operators. In this method, all the individuals left who after reproduction are taken for crossover operation and the individuals who cannot produce offspring with better fitness values from their fitness are changed[12].

V. Dheepa et al have presented the method for detecting the credit card fraud using SVM, it is an emerging trend for solving the classification related problem[13]. Venkata ratnam et al has proposed outlier detection algorithm in data stream using k-nearest neighbours. This outlier specifies the abnormal transaction characteristics and the transaction incidence[17]. Shailesh S have proposed a Neural network is used for classifying the transaction behaviour as special clusters from low to very high risk, uses SMNNN method for classifying into different group of clusters [14]. Rong-Chang Chen et al proposed a modified method for fraud identification using SVM and Neural Networks specially BPN (Back Propagation Network) in order to handle the unusual behaviours of credit card holder [21]. Pre-processing of data received in transaction level in the form of data aggregation allows to hold heterogeneous multidimensional data efficiently by taking into account the essential fields to develop aggregate model for identifying suspicious transactions [22]. The way business and financial institutions take more attempt to prevent fraud is effective when the two main issues cost savings and time efficiency are considered. Renuga devi et al have proposed a transaction based fraud detection. They have tracked by the behavioral sample of the customer by implementing the Classifiers Random Forest and Naïve Bayesian to predict the legitimate and fraudulent patterns. Also recommended that the aggregate model achieves better than personalized model. Naïve Bayesian approach attains best results for personalized model and Random Forest attains best results[23]. Surbhi Agarwal et al has proposed a fast fraud detection using hybrid approach such as clustering and outlier detection. Using clustering data sets are partitioned and outlier detection is used to find the fraudulent data. They have suggested that their two techniques are combined to efficiently find the outlier from the data set. Amruta D has presented a review on outlier detection techniques for credit card fraud detection. They have discussed one such unsupervised method Principal Component Analysis(PCA) to detect an outlier[24].

Pandey et al [20] have proposed a Decision Tree with Reduced Error Pruning Method and achieved 75.73% classification accuracy. Bashir et al achieved a classification accuracy 81.82% with combination of Naïve Bayes, Decision Tree and Support Vector Machine. Chaurasia et al [21] have used the commonest types of decision tree algorithms for the prediction of heart diseases. CARD, ID3 and DT decision trees were applied with the same dataset available at [10], and evaluated using 10-fold cross validation method. CARD decision tree has presented the highest classification accuracy with 83.49%, followed by DT with 82.50% and finally 72.93% for ID3. Uppin et al. have proposed C4.5 decision tree classifier for predicting heart disease. The data set in [13] were also used within this experiment. Their strategy aimed to reduce the number of parameters within the data set in order to avoid the redundant features that are not important in the classification. Therefore, 7 of 13 parameters have only been used and the C4.5 classifier showed 85.96% of accuracy. Mahmood and Kuppa has proposed a new pruning method with the aim of improving classification accuracy of heart diseases and reducing the tree size. A combination of pre-pruning and post-pruning was used for pruning C4.5 decision tree classifier. The new decision tree has been compared with the benchmark algorithms using dataset available online at [10]. The results showed that the new method significantly reduced the tree size and achieved 76.51% of accuracy. Shouman et al [24] have focused on the improvement of decision tree accuracy for diagnosis of heart disease. K-means clustering was integrated into the decision tree in order to enhance the diagnosis of heart disease. The dataset mentioned in [10] has been utilised. The highest accuracy obtained was 83.9% by applying the inlier method with two clusters. Melillo et al [21] developed a model for risk assessment in patients suffering from congestive heart failure. ECG recording for long-term heart rate variability has been used as a dataset, which is derived from two different Congestive Heart Failure databases. A CART decision tree algorithm is used with the aim of classifying patients into two groups based on the risk factor, and achieved 85.4% of accuracy. Bohacić [25] applied an alternating decision tree for the prediction of heart failure and obtained 77.65% classification accuracy. Nguyen, Abbas, Douglas, and Saeid (2015a) presented a medical diagnosis system which was mutual genetic fuzzy logic system with wavelet. The wavelet transformation was engaged to extract discriminative patterns for high-dimensional datasets from UCI. Then fuzzy standard additive trained by genetic algorithm (GSAM) was used to classifier medical dataset. This proposed method was evaluated using Cleveland heart disease datasets from UCI. The experimental results proved that GSAM became highly capable when deployed with small number of wavelet patterns as its computational load was reduced. However, this proposed approach had a shortcoming regarding selection of the best number of wavelet features and the accuracy of this proposed model was 78.78% for Cleveland heart disease datasets.

METHODS AND MATERIALS

Clustering:

Clustering is the process of grouping similar objects that are different from other objects. Clustering is an unsupervised classification technique, which means that it does not have any prior knowledge of its data and
results before classifying the data. The term clustering is also used by several research communities to describe the method of grouping unlabeled data. Clustering is used to improve the efficiency of the result by making groups of the data. So to cluster the data means specifying the data objects to a specific cluster which has similar objects or a group of objects.

**Farthest First:**

Farthest first is a Variant of K means that places each cluster centre in turn at the point farthest from the existing cluster centers. This point must lie within the data area. The objects that are farther are clustered simultaneously. This feature of farthest first clustering algorithm speeds up the clustering process in many situations like less relocation and less alteration is needed.

**Classification:**

Classification is a data mining (machine learning) method used to expect group membership for data instances. For example, you may desire to use classification to forecast whether the weather on a specific day will be “sunny”, “rainy” or “cloudy”. It is a mutual process related to classification, the process in which ideas and objects are recognized, differentiated, and understood. Classification is a common process related to categorization, the process in which ideas and objects are predictable, discriminated, and implicit.

**Decision Tree:**

Larger programs are usually split into more than one class. A decision tree is a predictive machine-learning method that decides the target value (dependent variable) of a new sample based on various attribute values of the available data. The interior nodes of a decision tree denote the different attributes, the branches between the nodes notify that the possible values that these attributes can have in the observed samples, while the terminal nodes notify that the final value (classification) of the dependent variable. The attribute that is to be calculated is known as the dependent variable, since its value depends upon, or is determined by, the values of all the other attributes. The other attributes, which help in calculating the value of dependent variable, are known as the independent variables in the dataset.

**Decision Tree : J48 Algorithm:**

Decision tree J48 is the implementation of algorithm ID3 (Iterative Dichotomiser 3) developed by the WEKA project team. It is an open source Java implementation of the C4.5 algorithm in the Weka data mining tool. The Decision tree J48 classifier provides the following simple steps... In order to classify a new item, it first needs to create a decision tree based on the characteristic values of the available training data. So, whenever it encounters a set of items it classifies the attribute that differentiates the various instances most clearly. This feature that is clever to tell us most about the data request so that we can classify them the best is said to contain the highest information gain. Now, among the achievable values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its group have the same value for the target variable, then we stop that branch and assign to it the target value that we have attained. For the other cases, we then look for another attribute that gives us the highest information gain. Hence we carry on in this way until we either get a clear decision of what arrangement of attributes gives us a particular target value, or we run out of attributes. In the event that we run out of attributes, or if we cannot get an unambiguous answer from the available information, we give this branch a target value that the majority of the substances under this branch possess. By checking all the respective attributes and their values with those seen in the decision tree representation, we can assign or predict the target value of this new instance.

**Naïve Byes:**

The NB algorithm is based on Bayes’ theorem and is a well known and simple probabilistic classifier widely used for real data sets. It is fairly insensitive to the number of features used and can be applied to all types of features. The design of NB presumes that all features are autonomous of one another. If redundant features are present or there are dependencies among features in the data set, the results can be confusing. The result of reducing irrelevant or redundant features shown up in better performance of the NB classifier.

**Iqr[Inter Quartile Range]:**

An outlier filtering approach commonly uses a distance measures to detect outlier instances that are at a significant distance from the others. It is a challenging task to eliminate outliers in order to improve the performance of classifiers. In this paper, Inter-Quartile Range to detect Outliers and Extreme Values are used in data sets.

Consider the following data points 32,26,27,11.6,28.5,33.2,18.9,48,41.2,25,36.1,24.6.

In step 1, order the sample data points in ascending order. The order of the data points follows the sequence.
In step 2 and 3, the first and third quartile values are calculated. The value of IQR is computed by following

\[
Q_1 = 24.6 \\
Q_2 = 27 \\
Q_3 = 33.2 \\
Q_4 = 48
\]

In step 4, the inter quartile range (IQR) is calculated by using Eq(1)

\[
\text{IQR} = Q_3 - Q_1 = 33.2 - 24.6 = 8.6. \\
\text{Where} \quad Q_3 = 33.2 \\
Q_1 = 24.6
\]

The lower boundary value is calculated using Eq(2).

\[
\text{Lower boundary} = Q_1 - (1.5 \times \text{IQR}) = 24.6 - (1.5 \times 8.6) = 24.6 - 12.9 = 11.7
\]

So the value of the lower boundary is 11.7. In step 4, the value of the upper boundary is calculated by using Eq(3).

\[
\text{Upper boundary} = Q_3 + (1.5 \times \text{IQR}) = 33.2 + (1.5 \times 8.6) = 33.2 + 12.9 = 46.1
\]

From the above lower and upper boundary values, the data points below and above the lower and upper boundary values should be considered as outliers. Here, data point 11.6 is below the lower boundary values so it should be an outlier point and the data point 48 is above the upper boundary value so it is also considered as an outlier point.

From the above steps, lower and upper boundary values can be calculated by using Eq(4) and Eq(5)

\[
\text{Upper Boundary} = Q_3 + (\text{EVF}\times\text{IQR}) \\
\text{Lower boundary} = Q_1 - (\text{EVF}\times\text{IQR})
\]

Where, EVF – Extreme Value Factor.

From the above upper and lower boundary calculations, the data points above and below are considered as extreme values. These extreme values are also considered as outliers.

**IV. Proposed Models:**

**A. Outlier Finding With Decision Tree (Ofwdt):**

In this model, a combined approach such as farthest first clustering, ODA and J48 classifier for diagnosing the breast cancer, Diagnosis of heart disease and credit card fraud detection is used. Farthest first clustering is used for clustering the original data set. Outlier detection algorithm (ODA) is used for detecting outliers. After the outliers have been removed, they are given as input into a J48 classifier to categorize the data into different classes. A variety of learning algorithms can be used for classification. Amongst the most commonly applied classifiers are the J48 algorithm (John & Langley, 1995). The following steps are used for OFWDT

1. Apply Farthest first clustering for preprocessing the original data set
2. Apply Outlier Detection algorithm on preprocessed data by following steps
   - **Step a:** Clusters having fewer numbers of objects:
     If a cluster using Farthest first contains only fewer numbers of points than the required number of outliers, the radius pruning is avoided for that cluster.
   - **Step b:** Pruning objects inside each cluster:
     Calculate distance of each point of a cluster from the radius of the cluster. If the distance of a point is less than the radius of a cluster, the point is pruned.
   - **Step c:** Detecting outlier objects:
     Calculate LOF for all the points that are left unpruned in all the clusters. If the outlier factor is greater than threshold then it will declare as outlier otherwise it is not a outlier. The Local outlier factor of \( p \) can be calculated by using Eq(1)
3. Remove outliers values from the original Data set
4. J48 classification algorithm applied in pre-processed data using WEKA
5. Classify the data with better accuracy using results of J48.

B. Outlier Finding With Naïve Bayes (Ofwnb):
In this model we used combined approach such as RODCA and NB classifier for detection of credit card fraud. Initially outlier and extreme values has been removed from the original dataset using RODC. After the outliers have been removed, they are input into a naïve Bayes classifier to categorize the data in to different classes.

Rodca [Robust Outlier Detection Using Clustering Approach]:
The proposed RODCA consist of two phases. In first phase all the data objects are initially clustered using farthest first. Every cluster is checked for number of objects it contains. If a cluster has less than average number of objects (n/k) then all the substances in the cluster are detected as outliers. That is, the whole cluster is selected as outlier cluster. In the second phase, the remaining outliers are detected by using following. For each cluster the absolute distance between each object and the cluster centroid is calculated. This is called the Absolute Distance between the centroids and Point (ADCP) [9] distance. For each cluster a threshold rate is also calculated as follows:

\[
\text{Threshold (T)} = (\text{Average ADCP}) \times 1.5
\]

Threshold is the lowest possible input value of similarity necessary to join two objects in one group. The value of T is calculated as the average of all ADCP values of the equal cluster multiplied by 1.5. Henceforth each ADCP value is compared with the corresponding value of threshold of its cluster. If the ADCP value is less than the threshold value the object is maintained in the cluster, but, if the ADCP value is greater than the threshold value the object is detected as outlier. The algorithm for robust outlier detection using clustering approach[RODCA] is outlined in Table 3. The detailed step for OFWNB Model is follows.

1. Pre-process data for applying RODCA technique using WEKA Tool.
   a) Perform Farthest First (FF) clustering algorithm to produce a set of k clusters.
   b) Determine small clusters and consider the points (objects) that belong to these clusters as Outliers
   For the rest of the clusters not determined in Step a:
      Begin
      c) For each cluster, compute the ADCP and T values.
      d) For each object in cluster, if ADCP > T then classify that object as an outlier; otherwise Not.
      e) Remove outliers and extreme values from the original Data set
2. Pre-processed dataset uploaded in WEKA toolkit for analysis.
3. Naïve bayes classification algorithm applied in pre-processed data using WEKA
4. Classify the data with better accuracy using results of Naïve Bayes.

C. Outlier Finding With Quartile Range (Ofwqr):
In this model, a hybrid technique is used. First, inter quartile filtering approach is implemented. The outliers in a data set are detected by using IQR. After pre-process the data, the decision tree J48 algorithm is executed on pre-processed data instances. The detailed steps for OFWQR model is follows

STEP 1: Pre-process data for applying Inter Quartile Range filter
a) Calculate the first and third quartile value (Q1, Q2)
   b) calculate Lower boundary = Q1 - (1.5*IQR)
   c) calculate Upper boundary = Q3 + (1.5*IQR)
   d) Data points anything outside the lower and upper boundary value is an outlier.

STEP 2: Calculate extreme values by following steps
a) Find the extreme value factor (EVF)
   b) Calculate Lower boundary = Q1 - (EVF*IQR)
   c) Calculate Upper boundary = Q3 + (EVF*IQR)
   d) Any data points outside the lower and upper boundary value are extreme values

STEP 3: Decision tree J48 classification algorithm applied in pre-processed data using WEKA
STEP 4: Classify the data with better accuracy using decision tree.
V. Experimental Study And Results:

Experimental Design:

To evaluate outlier detection, we ran our outlier models on three real life data sets (WBCD, Cleveland heart disease and German credit card) obtained from the UCI Machine Learning Repository. We have proposed three different methods for outlier detection using data mining algorithms. In this section, we compare the performance of the three outlier methods on identifying true outliers with three different data sets.

A. F-Measure:

F-measure is the choral mean of precision and recall, as given by Eq. (2). F-measure takes on values from 0 to 1. The top value of F-measure, the better the classification algorithm is.

\[ F \text{ measure} = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \]  

(2)

B. Precision:

Precision is the proportion of the true positives against all the positive results (both positive tuples and negative tuples), as given by Eq(3)

\[ \text{Precision} = \frac{TP}{TP + FP} \]  

(3)

C. Recall:

Recall is the proportion of the positive tuples against positive tuples and negative tuples, as given by Eq(4).

\[ \text{Recall} = \frac{TP}{TP + FN} \]  

(4)

D. Classification Accuracy:

Classification accuracy (ACC) is one of the most accepted metrics in classifier evaluation. It is the proportion of the number of positive tuples and negative tuples obtained by the classification algorithms in the total number of instances, as given by Eq.(5)

\[ \text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \]  

(5)

E. Roc Area:

It is another common metric for evaluating the classifiers. It equals to the possibility that a classifier will rank an arbitrarily chosen positive instance higher than an arbitrarily chosen negative one. It takes on values from 0 to 1. The better classification is based on the higher the value of ROC.

F. Kappa Statistics:

Kappa statistics is another evaluation metric which is based on the difference between the actual concurrence in the error matrix and the chance concurrence. The values for Kappa range from 0 to 1 and a perfect classification would make a Kappa value of 1. The equation for Kappa statistics is given in Eq. (6).

\[ \text{kappa} = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \]  

(6)

Where Pr(a) denotes the relative observed concurrence among raters and Pr(e) is the hypothetical probability of chance agreement

Results On Wisconsin Breast Cancer Data(Wbcd):

The first is the WBCD data set, which can be found in the UCI Machine Learning Repository. The data set contains 699 instances with 9 continuous attributes (not including the class attribute). Each instance is labeled as benign (458 or 65.5%) or malignant (241 or 34.5%). Here we follow the experimental technique of Harkins et al. by removing some of the malignant instances to form a very unbalanced distribution. 120 extreme values have been removed (all are malignant) from the dataset to construct a new dataset with 578 instances. Here malignant instances are deemed as outliers. The performance measures of three outlier models on breast cancer Wisconsin data set is outlined in Table 3.

Table 3: The performance measures of three outlier models on Wisconsin Breast Cancer Data set

<table>
<thead>
<tr>
<th>Outlier Model</th>
<th>Acc (%)</th>
<th>Precision</th>
<th>F-measure</th>
<th>Recall</th>
<th>Kappa Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFWDT</td>
<td>99.9</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>OFWNB</td>
<td>95.8</td>
<td>0.96</td>
<td>0.959</td>
<td>0.959</td>
<td>0.909</td>
</tr>
<tr>
<td>OFWQR</td>
<td>94.8</td>
<td>0.949</td>
<td>0.949</td>
<td>0.948</td>
<td>0.886</td>
</tr>
</tbody>
</table>

From Table 3, we can see that although the classification of the accuracies of outlier finding with decision tree model(OFWDT) is higher than those of outlier finding with naïve bayes(OFWNB) and outlier finding with quartile range(OFWQR) models. The average accuracy of of OFWDT model is markedly higher than those of
them. Therefore, we can further conclude that OFWDT model performs better than all other methods on the cancer data set. The OFWDT obtained the 99.9% classification accuracy on WBCD data set.

Results On German Credit Card Data:

To scrutinize outlier detection for credit card fraud we use standard German Credit Card Fraud dataset which is available on UCI Machine Learning Repository. This dataset consist of 20 attributes and 1000 instances. The attributes are both numerical and categorical. The original german data set having 25 outliers. These outliers are detected by using outlier models. In Fig. 1 shows that the sample outliers detected on credit german dataset. We used 10-fold cross validation method for experimental analysis. we have partitioned the data set randomly in to 10 equal sized partitions. Then we used one of the partition to test, whereas rest of the partitions is devoted to train the base learner. The above procedure is repeated ten times so that each partition is used for test accurately one time. Then, a mean accuracy of the individual results is merged. The average results of the 10 fold cross validation results reported in this section. We have done the experiments on WEKA version 3.7.13. The outlier models are implemented with filtering package of WEKA. The naïve bayes algorithm is implemented with classification package of WEKA.

Fig. 1: Sample Outliers in Credit Card Data

The performance measures of three outlier models on breast cancer Wisconsin data set is outlined in Table 4.

Table 4: The performance measures of three outlier models on Cleveland Heart Disease Data Set

<table>
<thead>
<tr>
<th>Outlier Model</th>
<th>Acc (%)</th>
<th>Precision</th>
<th>F-measure</th>
<th>Recall</th>
<th>Kappa Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFWDT</td>
<td>99.8</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.959</td>
</tr>
<tr>
<td>OFWNB</td>
<td>97.6</td>
<td>0.987</td>
<td>0.98</td>
<td>0.976</td>
<td>0.6554</td>
</tr>
<tr>
<td>OFWQR</td>
<td>96.3</td>
<td>0.927</td>
<td>0.945</td>
<td>0.963</td>
<td>0</td>
</tr>
</tbody>
</table>

From Table 4, we can see that although the classification of the accuracies of outlier finding with decision tree model(OFWDT)is higher than those of outlier finding with naïve bayes(OFWNB) and outlier finding with quartile range(OFWQR) models. The average accuracy of of OFWDT model is markedly higher than those of them. Therefore, we can further conclude that OFWDT model performs better than all other methods on the cancer data set. The OFWDT obtained the 99.9% classification accuracy on credit data set.

Results On Cleveland Heart Disease Data:

The Cleveland Clinic Foundation heart disease dataset has been used in this study, which is available online at [10]. It consists of 303 constant instances and without missing values. Each instance contains 14 attributes in addition to the output class, 54.5% of the instances are for patients with no risk of developing a heart failure, while the remaining 45.5% are for patients with different risk levels. In order to examine the overall performance of the decision tree classifier, the 10-fold cross-validation method is used to measure the classifiers' performance. This method is usually utilized to maximize the use of the data set. The data set is arbitrarily partitioned into 10 equal subsets. Each one of them contains approximately the same proportion of different class labels. Of the 10 subsets, a single subset is retained as a testing data, and the remaining 9 subsets as the training data.

The cross-validation method is then repeated 10 times, until each one of the 10 subsets was used accurately once as a testing set. The results can then be averaged to estimate the classifier’s performance. The advantage of
this model is that all subsets are used for both training and testing, and each subset is used for testing exactly once [13, 14].

The performance measures of three outlier models on german credit card data set is outlined in Table 5.

**Table 5: The performance measures of three outlier models on german credit card data.**

<table>
<thead>
<tr>
<th>Outlier Model</th>
<th>Acc (%)</th>
<th>Precision</th>
<th>F-measure</th>
<th>Recall</th>
<th>Kappa Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>OFWDT</td>
<td>99.9</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
<td>0.998</td>
</tr>
<tr>
<td>OFWNB</td>
<td>86.1</td>
<td>0.869</td>
<td>0.864</td>
<td>0.861</td>
<td>0.863</td>
</tr>
<tr>
<td>OFWQR</td>
<td>99.67</td>
<td>0.993</td>
<td>0.997</td>
<td>0.995</td>
<td>0.0</td>
</tr>
</tbody>
</table>

From Table 5, we can see that although the classification of the accuracies of outlier finding with decision tree model(OFWDT) is higher than those of outlier finding with naïve bayes(OFWNB) and outlier finding with quartile range(OFWQR) models. The average accuracy of OFWDT model is markedly higher than those of them. Therefore, we can further conclude that OFWDT model performs better than all other methods on the cancer data set. The OFWDT obtained the 99.8% classification accuracy on credit data set. From Tables 3–5, we can see that OFWDT model has the best performance. OFWDT model always the highest classification accuracy for the three data sets among all outlier detection methods. The classification accuracy of proposed outlier models on three data sets are outlined in Table 6.

**Table 6: Classification accuracies of three data sets**

<table>
<thead>
<tr>
<th>Data Sets</th>
<th>OFWDT</th>
<th>OFWNB</th>
<th>OFWQR</th>
</tr>
</thead>
<tbody>
<tr>
<td>WBCD</td>
<td>99.9%</td>
<td>95.8%</td>
<td>94.8%</td>
</tr>
<tr>
<td>German Credit</td>
<td>99.8</td>
<td>97.6%</td>
<td>96.3</td>
</tr>
<tr>
<td>Cleveland Heart</td>
<td>99.9%</td>
<td>86.1</td>
<td>99.67%</td>
</tr>
</tbody>
</table>

From Table 6 we can clearly see that outlier finding with decision tree model performs better based on the classification accuracy obtained on three data sets. The comparison results three data sets based on classification accuracy is outlined in Table 7. The methods of fraud detection on credit card data set is outlined in Table 8.
From Table 7 we can see that the proposed model obtained better classification accuracy on breast cancer and heart data set compared with existing research on diagnosis of breast cancer and heart disease. From Table 8 we see that the numbers of methods are used for fraud detection. Among all the existing methods, the proposed model give better sensitivity and specificity ie. True positive and False positive.

Discussion:

Roughly speaking, the current methods to outlier detection can be classified into the following five categories.

(1) Distribution-based method is the classical method in statistics. It is based on some standard distribution model (Normal, Poisson, etc.) and those objects which deviate from the model are recognized as outliers. Its greatest disadvantage is that the distribution of the measurement data is unknown in practice. Since our method does not require knowledge about the distribution of the measurement data, it can counter the main limitations of the distribution-based method.

(2) Depth-based method is based on computational geometry and compute different layers of k-d convex hulls and flags objects in the outer layer as outliers. However, it is a well-known fact that the algorithms employed suffer from the dimensionality curse and cannot cope with large k. Comparing with the depth-based method, the time complexity of our method is relatively low, which is more suitable to deal with large data sets.

(3) Clustering-based method classifies the input data. It detects outliers as by-products. However, since the main objective is clustering, it is not optimized for outlier detection. Differing from the clustering-based method, our method is designed specially for outlier detection.

(4) Distance-based method was originally proposed by Knorr and Ng (1998), Knorr et al. (2000). In traditional distance-based outlier detection, being an outlier is regarded as a binary property, we only know that an object is an outlier or not. In our method, we introduce a notion called local outlier factor, which can indicate the degree of outliers for each object. Moreover, although our method also calculates the distances between objects, we adopt different attitudes to objects from different parts of the given data set when detecting outliers based on distance.

(5) Density-based method was originally proposed by Breunig et al. (2000). A local outlier factor (LOF) is assigned to each sample based on their local neighborhood density. Samples with high LOF value are identified as outliers. The disadvantage of this solution is that it is very sensitive to parameters defining the neighborhood. Unlike density-based outlier detection, our method does not require such parameters. And our method can also find the local outliers as the density-based method does, since the definition for outliers in our method has a characteristic that is ignored by most current definitions for outliers. That is, for a given data set, we consider detecting outliers with respect to any subset of the given data set, which makes it possible to find the local outliers.

From the experiment results in Section V, we can see that the performance of OFWDT method can effectively improve the performance of OFWNB, OFWQR models. On the other hand, our method can also solve the main problems of distance-based method.

Conclusion:

Outlier detection is becoming critically important in many areas. In this paper, we presented a new method for outlier detection. Experimental results on real data sets demonstrated the effectiveness of our method for outlier detection. The performance of OFWDT method is better than those of OFWNB and OFWQR methods. This indicates we obtained a more effective method for outlier detection based on the classification accuracy.