International Journal of Sciences: Basic and Applied Research (IJSBAR)

ISSN 2307-4531
(Print & Online)

http://gssrr.org/index.php?journal=JournalOfBasicAndApplied

A Hybrid Ensemble Method for Multiclass Classification and Outlier Detection

Dalton Ndirangu[a*], Waweru Mwangi[b], Lawrence Nderu[c]

[a]*Lecturer, United States International University-Africa, P.O. Box 14634 00800, Nairobi, Kenya*

[b]*Professor, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62,000 – 00200 Nairobi, Kenya*

[c]*Lecturer, Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62,000 – 00200 Nairobi, Kenya*

[a]*Email: dndirangu@usiu.ac.ke*

[b]*Email: waweru_mwangi@icsit.jkuat.ac.ke*

[c]*Email: lnderu@jkuat.ac.ke*

Abstract

Multiclass problem has continued to be an active research area due to the challenges paused by the issue of imbalance datasets and lack of a unifying classification algorithms. Real world problems are of multiclass nature with skewed representations. The study focused on the challenges of multiclass classification. Multiclass datasets were adopted from UCI machine learning repository. The research developed a heterogeneous ensemble model for multiclass classification and outlier detection that combined several strategies and ensemble techniques. Preprocessing involved filtering global outliers and resampling datasets using synthetic minority oversampling technique (SMOTE) algorithm. Datasets binarization was done using OnevsOne decomposing technique. Heterogeneous ensemble model was constructed using adaboost, random subspace algorithms and random forest as the base classifier. The classifiers built were combined using average of probabilities voting rule and evaluated using 10 fold stratified cross validation. The model showed better performance in terms of outlier detection and classification prediction for multiclass problem. The model outperformed other commonly used classical algorithms. The study findings established proper preprocessing and decomposing multiclass results in an improved performance of minority outlier classes while safe guarding integrity of the majority classes.

Keywords: Multiclass; Outlier; Classification; Classifiers; Ensemble.

---

* Corresponding author.

1. Introduction

Multiclass problem has continued to be an active research are due to the challenges paused by the issue of imbalance datasets and lack of a unifying classification algorithms. Real world problems are of multiclass nature with skewed representations. According to [1] decomposition strategies have been demonstrated to be a successful methodology for multiclass classification problems. Multiclass datasets have might contain outliers associated with the features of the datasets or and classes with very few representations. The rare classes are also treated as outliers. Presence of outliers degrade performance of classifiers. Author [2] proclaimed that outlier detection is one of the important research tasks in data mining which essentially finds the observations that are deviating from the common expected behavior. The author concludes that there is no universally accepted methodology for detecting and analyzing outliers.

Data preprocessing is crucial in any data mining process. Accuracy performance of classification is improved when redundant and irrelevant features have been removed. Dimensional reduction of attributes is desirable because it reduces the complexity of the model resulting in a clear and understandable model. Performance of classifiers is greatly improved once the aspect of dimensional data reduction has been achieved through techniques such as feature selection and extraction [3]. According to [4], the ensemble technique can be applied to feature selection method. Ensemble feature technique involves use of different feature selection methods with the same training data or by use of single feature with different training data. The ensemble feature results from a combination of rankings of features that contain all the ordered features. The results of the base selectors are combined using different combination methods, and a practical subset is selected according to several different threshold values. The Ensemble techniques utilize the explicit power of multiple models to realize better prediction accuracy than the case when individual models are used. The ensemble learning algorithms used in the design should be competent enough and complementary to one another.

Study by [5] used information-gain, gain-ratio, chi-square and ReliefF filter methods to create an ensemble filter selection method. Their study showed combining feature selection methods improves the performance of classifiers by identifying the features that are weak as individual but strong as a group. This study aimed at developing a hybrid ensemble method for multiclass classification and outlier detection using adaboost, random subspace (RSM) algorithms and random forest (RF) as the base classifier. The proposed method incorporates several strategies and ensemble techniques.

The rest of the paper is organized as follows: Critique of existing classification and outlier literature is provided. Problem statement is derived from the critiqued literature. A proposed method is provided. Experiments and results are provided in form of tables and figures. Results are discussed in line with the literature and conclusion drawn from the discussions. Recommendations and references are provided at the end.

2. Critique and review of classification works

Authors [6], addressed the issue of imbalanced dataset by use of synthetic minority over-sampling technique (SMOTE) approach, which generate synthetic minority samples. The technique enabled improved visibility of

minority classes. However, their study was only based on binary classification problem. According to [7] classification systems play an important role in business decision-making tasks by classifying the available information based on some criteria. The objective of their research was to assess the relative performance of some well-known classification methods. Their research experimental findings affirmed that data characteristics considerably impact the classification performance.

Author [8] in his research on survey of classification techniques in data mining concludes that decision trees and bayesian network (BN) in most cases have significant different operational profiles in a way that, when one is very accurate the other is not and vice versa. On the contrary, author argued that decision trees and rule classifiers have a similar operational profile. Thus different algorithms performs differently and thus amalgamating algorithms through ensemble method produces a more robust method.

.Author [9] proposed a novel over sampling strategy to handle imbalanced data based on ensembles, named cluster ensembles based SMOTE (CE-SMOTE), which first used cluster ensemble to generate multiple partitions. However, their study focused only on binary classification. Most of the ensemble methods use a single base learning algorithm to produce homogeneous base learners. However there are also some methods that use learners of different types leading to heterogeneous ensembles. In order for ensemble methods to be more accurate than any of its individual members, the base learners have to be as accurate as possible and as diverse as possible [10].

Study by [11], on multiclass imbalance problems, showed that class imbalance problems have attracted much research focus due to classification difficulty caused by the imbalanced class distributions. They reaffirmed that many ensemble methods have been proposed to deal with such imbalance classes though a lot of earlier research focused on binary imbalance problems. They proclaimed that there are unsolved issues in multiclass imbalance problems, which exist in real-world applications. Their studies focused on the challenges posed by multiclass imbalance problems and investigated the generalization ability of some ensemble solutions. The study findings showed poor classification performance after applying multi-minority and multi-majority on their experimental dataset. However performance improved after applying AdaBoost.NC learning algorithm. The results suggest use of ensemble learning algorithms improves performance of classifiers.

Author [12] in a survey of efficient classifiers for multiclass classification problems maintained that classification problems have become more complex and intricate in modern applications in the face of continuous data explosion. The author showed their model can significantly improve classification training time by combining a compact subset of relevant features without the loss of accuracy in multiclass classification problems. In addition, the discrimination degree of their classifier outperforms other conventional classifiers. The study indicates that multiclass classification problems continue to be a problem and more research is required in this area of multiclass classification.

There have been major changes and evolution done on classification of data [13]. Author noted that as the application area of technology increases, the size of data also increases. Classification of data becomes difficult because of unbounded size and imbalance nature of data. Thus class imbalance problem becomes one of the

greatest issues in data mining. They reaffirmed that feature selection method can also be used for classification of imbalance data.

In [14], proposed a unifying framework for multiple classifier systems that conceptually unifies a large variety of ensemble classification methods, including existing class binarization techniques such as error correcting output code (ECOC). Binarization is a technique of extending binary classification algorithm to multiclass problems through the process of decomposing the original multiclass problem into a series of smaller two class problems. Author concludes by proclaiming that there is relatively few theoretical studies that integrate the strategy of data manipulation approaches with the learners' manipulation approach. Study by [15] on ensemble learning classification algorithm proposed a novel feature selection method that improves performance of classifiers. Author noted that ensemble learning method improves the classification performance of single classifier. Their experimental findings showed that feature selection method based on discrimination and class information of each feature ensemble classifiers can achieve higher predictive accuracy than several classical feature selection methods.

In another study on improving performance on ensemble classifiers, [16] studied two different feature selection methods using Colon Cancer dataset. Different ensemble methods were implemented to the dataset having reduced features. Performance improvements obtained by this method were evaluated using the individual and ensemble classification methods. Experimental findings reviewed that classification accuracy for the colon cancer can be increased by use of both feature reduction and ensemble methods.

Many ensemble methods, seek to promote diversity among the models they combine [17]. Author [18] reaffirmed that RSM is one of the ensemble learning algorithms widely used in pattern classification applications. They reiterated that RSM has the advantages of small error rate and improved noise insensitivity due to ensemble construction of the base-learners. However, random selection of feature subspaces may result in a poorly selected feature subsets leading to poor discrimination capability. The latter cause a reduction of the final ensemble decision performance because of contributions of classifiers trained by subsets with low class separability. They concluded that the technique of vote weighting may overcome the drawbacks of RMS. Author [19] studied SMOTE-based classification approach to online data imbalance problem. Their results findings demonstrated that the SMOTE improves on the generalization performance of classifiers. The Random Subspace Method (RSM) introduced by [20] is an attractive choice for classification problems. Similar to bagging, it benefits from bootstrapping and aggregation.

3. Critique and review work on outliers detection

Although there are a number of methods for detecting outliers in a given dataset, no single method is found to be the universal choice. Depending on the nature of target application, different applications require use of different detection methods. There is need to develop new outlier detection method using either data centered or algorithmic approach [21]. Author [22] assert that outlier detection is an important research area forming part of many application domains. Their survey tried to provide a structured and comprehensive overview of the research on classification based outlier detection. They proclaimed that current research done on outlier

detection is in an unstructured fashion. They proposed a possible future work to unify the assumptions made by different techniques regarding the normal and outlier behavior into a statistical or machine learning framework.

In another study, Reference [23] affirms that outlier detection is an important research problem in data mining that aims at discovering useful abnormal and irregular patterns hidden in large datasets. Author argues that outlier detection has become the enabling underlying technology for a wide range of practical applications in industry, business, security and engineering, etc. According to [24], combination of ensemble learning with resampling techniques produce better classifiers that outperform other individual classic algorithms. Author [25] assert that rare class classification is the data mining technique for building a model that can correctly classify both the majority and minority (outlier) classes. Classifying minority or rare class is difficult because size of the rare class representation is too small and existing classification algorithms were designed to be biased towards prediction of majority class.

Study by [2] on comparative analysis of outlier detection techniques reviewed that outlier detection is one of the important aspects of data mining which essentially finds observations that are deviating from the common expected behavior. In their study, they provided a broad and a comprehensive literature survey of outliers and outlier detection techniques. Their findings reaffirmed that there is no universally accepted scale of any methodology to detect and analyze outliers. In a review study on outlier, [26] affirms that with advances in hardware and software technology, there has been a large body of work on temporal outlier detection from a computational perspective within the computer science community. The advanced technologies have accelerated and promoted the growth of different kinds of datasets such as data streams, spatio-temporal data, distributed streams, temporal networks, and time series data, generated by a multitude of applications. The authors further suggested the need for an organized and detailed study of the work done in the area of outlier detection with respect to such temporal datasets. They concluded by affirming that the methods for different data types are not easy to generalize to one another, though some of them may have similarity in the framework at the broader level.

In another survey on outlier detection methods, Reference [27] reaffirmed that outlier detection is an active area for research in dataset mining community. They emphasized the need for creating more attention while analyzing outlier. Detecting outliers and analyzing large datasets can lead to discovery of unexpected knowledge in area such as fraud detection, telecommunication, web logs, and web document, etc. Their findings showed that most of the techniques used for outlier detection focus more on algorithms and ignores data. It is observed that efficiency of outlier detection method is highly dependent upon data distribution and type of data.

Author [28] carried study on automatic outlier identification in data mining using inter quartile range (IQR). They asserted that some of the real time databases contain exceptional values or extreme values generally referred to as outliers. They proclaimed that separation of outliers from dataset is very important as it leads to improvement of data quality. They further stated that outliers influence the results of data mining techniques like clustering, classification and association.

Depth based outliers on squared-well was developed by [29]. The approach is able to improve on time and

efficiency for detecting outliers in large dataset. It is arguably established that this algorithm of the detection of the outlier is much effective than the previous one of depth based. In an effort to solve problems of outlier detection in an imbalance data set , Reference [30] combined both local and global outlier detection and proposed a method which clearly handle data having imperfect labels and enhanced the performance of outlier detection. The approach has been applied in real life dataset and the experimental findings concludes that these proposed approaches attain better tradeoff between false alarm rate and detection rate as compared to the traditional techniques.

Author [31] surveyed different unsupervised techniques for outlier detection. They asserted that methods used for outlier detection are application specific. Moreover selection of outlier detection method also depends on the type of data involved. Different methods can be used to detect outliers. However, outlier detection can be efficient if one method is used as a compliment to other method, so that the drawbacks of one method are minimized by use of the other method. Author [32] suggested a technique in which two algorithms could be used for the detection of the outliers. They proposed the use of distance based outlier detection and cluster based outlier detection algorithm for the detection and removal of the outliers. These technique of the outlier detection can be used for various domains like big data, high dimensional data etc. Reference [33] refuted the claim that high dimensional data hinders the detection of outliers using distance base-methods. It has been widely perceived that as dimension of data increases, distance –based methods label all points almost equally as good outliers. Author in their research on reverse nearest neighbors in unsupervised distance-based outlier detection provided evidence supporting the opinion that such a view is too simple, by demonstrating that distance-based methods can produce more contrasting outlier scores in high-dimensional settings. Furthermore, they showed that high dimensionality can have a different impact, by reexamining the notion of reverse nearest neighbors in the unsupervised outlier-detection context. Previous studies had showed that the distribution of points' reverse-neighbor counts becomes skewed in high dimensions, resulting in the phenomenon known as hubness.

Author [34] constructed heterogeneous ensemble classifier using combination of bagging, boosting and random subspace ensemble algorithms. Their choice was driven by knowledge that bagging, boosting and random subspace methods are well known re-sampling ensemble methods that generate and combine a diversity of learners using the same learning algorithm for the base learners. The author used model tree inducer M5rules in one set of experiment and SMO algorithm in another set of experiment. They compare the performance of the model with the individual learning algorithms used in the construction of the heterogeneous ensemble classifier. However author did not balance the classes nor attempted to remove point outliers that could be associated with some attributes of some of the datasets.

Author [35] researched on hybrid outlier detection method for Health Care Big Data. They reiterated that informatics, digitalizing health records, and telemedicine has resulted in rapid growth of health care data. Author noted that the conventional outlier detection methods are sometime not very efficient. In their study, they proposed a novel hybrid outlier detection method, namely, pruning-based K-nearest neighbor (PB-KNN), which integrates the density-based, cluster-based methods and K-nearest neighbor algorithm (KNN. Their findings showed that the PB-KNN method outperforms the k-nearest neighbor (KNN) and local outlier factor (LOF) in terms of the accuracy and efficiency because of effectively reducing and pruning data dimensionality.

Study by [36] reaffirmed that outlier detection is one of the major issues in data mining especially in the area of pattern classification. They asserted that outliers that were initially considered as noise can no longer be ignored without being analyzed. The discovery of outlier is extremely useful in detection of unpredicted and unidentified data, in certain areas like business fraud detection, computer networks intrusion, criminology, drug manufacturing, medical diagnosis and others. Author [37] developed a model-based outlier detection system using IQR filter algorithm. They used the system to detect and remove point outliers. Their findings showed that removing point outlier improves on outlier detection and performance of classifiers.

4. Problem statement

Multiclass classification problem has many challenges which include, imbalance classes, presence of outliers in the features, redundant features and lack of robust learning algorithms. The literature review has shown that there is no general approach for solving multiclass problem and the unification framework does not exist. One of the directions of addressing imbalance classes which is a common phenomenon with multiclass classification problem is to study underlying nature of the imbalanced data, key properties of its underlying distribution and consequences they bring for learning better classifiers or for constructing specialized pre-processing methods. The review has shown there are a number of methods for detecting outliers in a given dataset but no single method is found to be the universal choice. Depending on the nature of target application, different applications require use of different outlier detection methods. Thus there is need to develop outlier detection method using either data centered or algorithmic approach. Rare class classification is the data mining technique for building a model that can correctly classify both the majority and minority (outlier) classes. Classifying minority or rare class is difficult because the size of the rare class representation is too small and existing classification algorithms were designed to be biased towards prediction of majority class. More research work is needed for the ever challenging emerging multiclass problem in real life applications. Classification and outlier share similar variant and bias and hence the need to combine the study of classification and outlier detection. Thus the study tend to answer the following questions:

(i) How does the methods for feature selection and outlier detection in data mining assist in improving performance in multi-class classification?

(ii) How can a model of multiclass classification and outlier detection be created using ensemble techniques?

(iii) How can test and evaluation of multiclass classification and outlier detection model be done using multiclass datasets?

Thus we propose development of a hybrid ensemble method for multiclass classification and outlier detection. The next section describe the proposed hybrid ensemble model for multiclass classification and outlier detection.

5. Proposed hybrid ensemble method

5.1 Ensemble Filter Feature Selection Method

The process use Correlation, Information gain, Relief, and Gain ratio filter feature selection algorithms. Figure 1 depict the process of developing the ensemble feature method. The process starts by selecting the four algorithms and using them to individually rank the features of the datasets. The generated feature lists are then merged together to produce a single ranked feature list.

Random forest classifiers are built using the features. The process begin by building classifier using the top ranked list. The Root Mean Square Error (RMSE) for the classifier is observed and recorded. The process is repeated iteratively by incorporating a feature at a time until the bottom ranked feature in the merged feature is used in the building of the classifiers. When a feature with less contribution to the performance of classifier is incorporated, the resulting classifier is expected to have a higher RMSE value compared with the previous immediate RMSE value. Thus the threshold is set to this level. The final optimal feature sub-list includes the features starting from the top-ranked feature up to and including the feature that results in the generation of the least RMSE value.

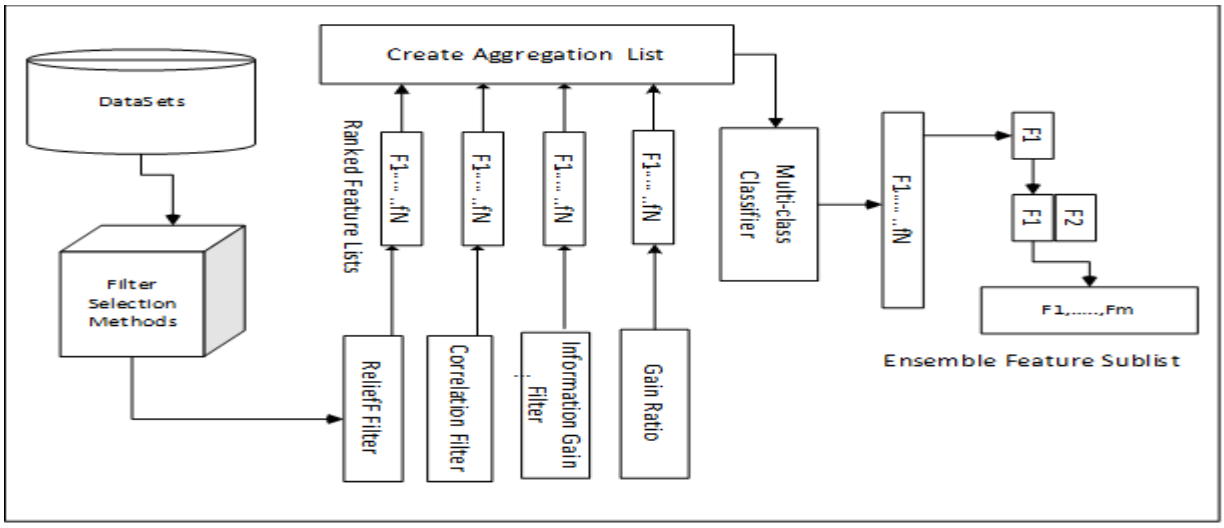

Figure 1: Ensemble Filter Selection Method

5.2 Preprocessing Dataset

Using the feature sub-list, point outliers are removed from the features using Inter-Quartile Range (IQR) algorithm. The outliers are identified from statistical tail ends as follows:

$$X \leq Q_3 + 3*IQR \text{ or } X \geq Q_1 - 3*IQR \quad (1)$$

Where: Q_1 = 25% quartile, Q_3 = 75% quartile and

IQR = difference between Q_1 and Q_3

Further preprocessing involve rebalancing datasets using Synthetic Minority Oversampling Technique (SMOTE). Artificial samples of the rare classes are generated to the level that they measured at least 50% compared with the majority classes.

5.3 Transforming Multiclass to Binary Classes

One-verses-One decomposition technique utilizing pairwise coupling is used to transform the preprocessed multiclass dataset classes to binary classes.

5.4 Building Ensemble Model

Using the decomposed dataset, heterogeneous ensemble model is built. The process involve generating two ensemble classifiers AD_RF and RS_RF using Adaboost algorithm and Random Subspace algorithm respectively each utilizing random forest algorithm as their base classifier. The two ensemble classifiers are then combined using voting technique utilizing average of probabilities combination rule. Figure 2 shows the proposed model.

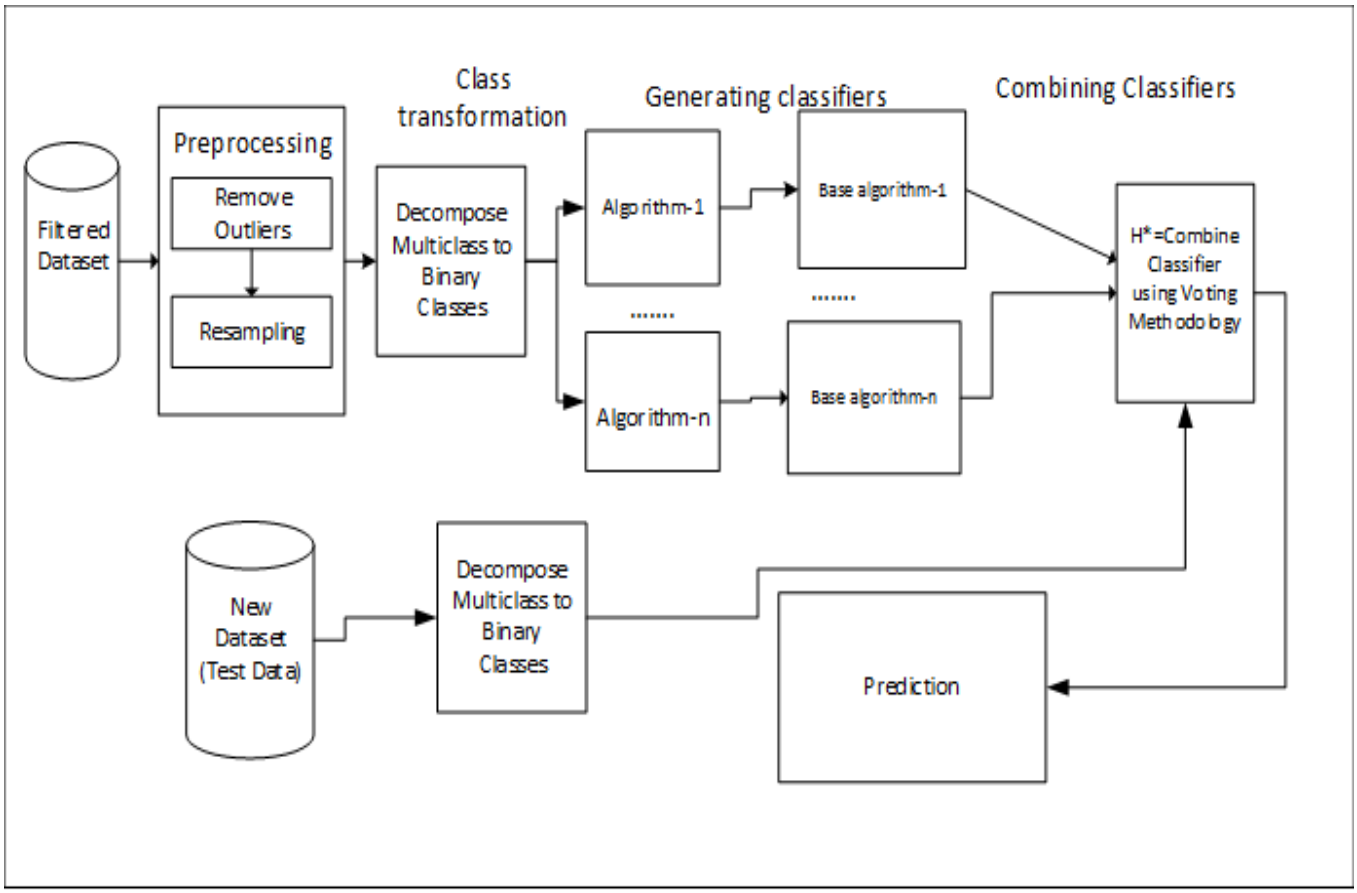

Figure 2: Proposed Ensemble Model

5.5 Testing and Validating Model

The model is validated using stratified 10 fold cross validation. Performance of the model is compared with other algorithms using a paired T-test. Statistical confidence interval is set at 0.05. Receiver Operating Characteristic (ROC) values is used to measure the performance of the classifiers. Other metrics performance measures such as True Positive, Precision, Recall and F-measure are used to evaluate the performance of the model.

6. Experiments and results

6.1 Dataset Descriptions

Since research was on multiclass classification and outlier detection, imbalance sensitive datasets were selected. The study used 6 multiclass medical datasets, 3 multiclass biological datasets and 1 manufacturing multiclass dataset drawn from UCI machine learning repository [38]. The description of the dataset is provided as follows:

(i) Cleveland Dataset

Cleveland dataset is a part of the Heart Disease Data Set (the part obtained from the V.A. Medical Center, Long Beach and Cleveland Clinic Foundation). The dataset was originally created to detect the presence of heart disease in the patient. It is integer valued from 0 to 4.The dataset 13 attributes, 5 classes and 297 instances.

(ii) Contraceptive Dataset

Contraceptive dataset is a subset of the 1987 National Indonesia Contraceptive Prevalence Survey. The samples are for married women who were either not pregnant or did not know if they were at the time of interview. The dataset was created to predict contraceptive method choice at the time of interview (no use, long-term methods, or short-term methods) for women based on their demographic and socio-economic characteristics. The dataset has 9 attributes, 3 classes and 1473 instances.

(iii) Dermatology Dataset

The dataset was original created to perform the differential diagnosis of erythemato-squamous diseases which is a real problem in dermatology. Patients were first evaluated clinically with 12 features. Afterwards, skin samples were taken for the evaluation of 22 histopathological features. The dataset has 34 attributes, 6 classes and 366 instances.

(iv) ***Ecoli Dataset***

The original Ecoli dataset is a multiclass classification dataset having 8 attributes. Here, 7 numerical attributes are utilized and the attribute “sequence name” is omitted. Among the 8 classes omL, imL, and imS are the minority classes and used as outliers. All the other majority classes are used as inliers. The dataset has 7 attributes, 8 classes and 336 instances.

(v) Glass Identification Dataset

The Original dataset was obtained from USA Forensic Science Service. The dataset has 6 types of glass which can be found in the crime scene, defined in terms of their oxide content (i.e. Na, Fe, K, etc). The dataset has 9 attributes, 7 classes and 214 instances.

(vi) Newthyroid Dataset

This dataset is one of the several databases about Thyroid available at the UCI repository. The dataset was created to detect whether a given patient is normal (1) or suffers from hyperthyroidism (2) or hypothyroidism (3). The dataset has 5 attributes, 3 classes and 215 instances.

(vii) Red Wine Quality Dataset

The dataset is related to red variant of the Portuguese Vinho Verde wine. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are much more normal wines than excellent or poor ones). The dataset has 11 attributes, 6 classes and 1599 instances.

(viii) Zoo Dataset

Zoo database was meant to classify animals in seven predefined classes and most of the attributes are boolean-valued. The dataset has 16 attributes, 7 classes and 101 instances.

(ix) Vehicle Dataset

Vehicle dataset is used to classify a given silhouette as one of four types of vehicle, using a set of features extracted from the silhouette. The vehicle may be viewed from one of many different angles. The dataset has 18 attributes, 4 classes and 946 instances.

(x) Yeast Dataset

This database contains information about a set of Yeast cells. The original use of dataset was to determine the localization site of each cell among 10 possible alternatives. The dataset has 8 attributes, 10 classes and 1484 instances.

7. Preprocessing datasets

An ensemble filter method was constructed using Correlation, Information-gain, Gain-ratio and ReliefF features selection algorithms and the resulting ranked lists merged. The features where thereafter evaluated and an optimal sub-list generated. Table 1 represents results.

Table 1: Feature Sublists for 10 Multiclass Datasets

	Datasets	#attributes	#instances without missing values	#classes	Selected Features	Dropped Features
1	Cleverland	13	297	5	3,8,9,10,11,12,13	1,4,5,6,7
2	Contraceptive	9	1473	3	1,2,3,4,5,6,8,9	7
3	Dermatory	34	358	6	2.3.4.5.9.14.15.17.20,21,22,26,27,28,31,33	1,6,7,8,10, 11, 12, 13,16, 18, 19,23, 24, 25, 29, 30, 32
4	Ecoli	7	336	8	1,2,3,5,6,7	4
5	Glass	9	214	6	1,2,3,4,6,7,8,9	5
6	Newthyroid	5	215	3	1,2,3,4,5	None
7	Redwine	11	1599	6	1,2,3,5,7,8,10, 11	4,6,9
8	Zoo	16	101	7	1,2,3,4,5,6,8,9,10,12,13,14,16	7, 11,15
9	Vehicle	18	946	4	1,2,3,4,5,6,7,8,9,10,11,12, 13,14,17,18	15,16
10	Yeast	8	1484	10	1,2,3,4,5,6,8	7

8. Effect of using ensemble filter method and point outlier detection

Point outliers were detected using IQR filter algorithm. Table 2 shows results of experiment before and after applying the developed ensemble feature selection method. We observe after applying the ensemble method, the number of detected point outliers reduced from 112 to 48 for Redwine dataset, 82 to 39 for Yeast dataset, 1 to 0 for Cleveland dataset. Other datasets showed no change since the filtered features did not contain any outlier.

Table 2: Effect of Ensemble Filter method on Point Outlier Detection

	Datasets	Detected Point-Outliers before Filtering	Detected Point-Outliers after Filtering
1	Cleverland	1	0
2	Contraceptive	1	1
3	Dermatory	0	0
4	Ecoli	0	0
5	Glass	16	16
6	Newthyroid	16	16
7	Redwine	112	48
8	Zoo	0	0
9	Vehicle	9	9
10	Yeast	82	39

9. Effect of removing point-outliers on classification

We sought to determine the effect of presence of point outliers on classification performance using the proposed method. Experiment was done using preprocessed Redwine dataset, one of the dataset used in the study. Table 3 shows results of experiment before and after removing point-outliers. Results indicate the overall weighted ROC classification performance of proposed method improved from 86.6% to 86.8%, SVM improved from 73.6% to 74.2%, SVM improved from 70% to 70.5%, KNN improved from 73% to 75.9%, OneR declined from 68.7% to 65.3%, C4.5 improved from 62.2% to 76.7% and randomforest improved from 72.9% to 86.1%. The proposed method had a better performance than other well known classification algorithms. Generally removing point outliers improved on classification performance of the classifiers.

Table 3: Effect of Removing Point-Outliers on Classification Performance

Algorithm	Weighted ROC Area (Outliers Removed)	Weighted ROC Area (with Outliers)
Naïve Bayes	0.742	0.736
SVM	0.705	0.7
KNN	0.759	0.73
OneR	0.653	0.687
C4.5	0.767	0.622
RandomForest	0.861	0.729
Proposed Ad_RF+ RS_RF	0.868	0.866

10. Comparison performance of proposed method with other algorithms using statistical paired-t test

Ten preprocessed multiclass datasets were used in the experiment. The performance of the proposed method was compared with the individual algorithms used to construct the method and also with other commonly used classification algorithms. ROC value was used as the metric performance measure. Performance evaluation was done using statistical Paired T-test with significant level p set at 95% confidence interval. Results were presented using some terms. The term "v" represents the winning situation of that particular algorithm as compared with the proposed algorithm while "*" indicate that the proposed ensemble algorithm was statistically better than the compared algorithm. **Plain text** signifies that there was no difference in performance indicating a draw. Aggregated results are represented in terms of x, y, and z where "x" represents number of aggregated losses and "z" represents aggregated number of wins and "y" represents aggregated number of draws for the proposed method.

10.1 Statistical Paired T-test between Proposed Method and other Algorithms before Applying SMOTE Resampling

Table 4 shows results of the experiment. The results indicate Ad_RF statistically lost once to the proposed

method. The proposed method had zero significantly error rate to RS_RF. Thus the propose method performed better than the individual classifiers used in the construction of the method. Further observation shows the proposed method outperformed Bagging by 20%, Naïve bayes by 30%, KNN by 70%, SVM by 90%, JRipper by 80%, OneR by 100%, ZeroR by 100% and C4.5 by 50%.

Table 4: Statistical Paired T-test ROC Performance between Proposed Method, Individual Classifiers and Other Classification Algorithms before SMOTE Resampling

Dataset	ProposedMethod	Bagging	Ad-RF	RF	RS_RF	NaiveBayes	KNN	SVM	Jripper	OneR	ZeroR	C4.5
'cleverland	0.89	0.88	0.88	0.88	0.9	0.9	0.78*	0.63*	0.56*	0.61*	0.5*	0.79*
contraceptiveDataset	0.73	0.76	0.71	0.74	0.72	0.7	0.63*	0.69*	0.66*	0.61*	0.5*	0.71
'dermatory	1	0.97	1	1	1	1	0.93*	0.88*	0.94*	0.5*	0.5*	0.95
ecolidataset	0.98	0.98	0.98	0.98	0.98	0.98	0.96	0.5*	0.94*	0.85*	0.5*	0.94
myglassdata	0.94	0.92	0.95	0.94	0.93	0.72*	0.8*	0.79*	0.82	0.74*	0.5*	0.8
newthyroid	0.99	0.97	0.99	0.99	0.99	0.99	0.97	0.6*	0.91*	0.88*	0.5*	0.9*
'RedWineQuality	0.89	0.85*	0.89*	0.89	0.89	0.78*	0.76*	0.7*	0.72*	0.69*	0.5*	0.76*
vehicle	1	0.98*	1	1	1	0.81*	0.96*	0.53*	0.94*	0.76*	0.5*	0.94*
yeastdataset	0.85	0.86	0.82	0.86	0.84	0.84	0.71*	0.71*	0.79*	0.49*	0.5*	0.73*
'Zoo	1	1	1	1	1	1	1	0.99	0.96	0.82*	0.5*	1
Average	0.93	0.92	0.92	0.93	0.93	0.87	0.85	0.7	0.82	0.69	0.5	0.85
	(x/ y/z)	(0/8/2)	(0/9/1)	(0/10/0)	(0/10/0)	(0/7/3)	(0/3/7)	(0/1/9)	(0/2/8)	(0/0/10)	(0/0/10)	(0/5/5)

10.2 Statistical T-test between Proposed Method and Other Algorithms after SMOTE Resampling

Table 5 shows result of experiment after SMOTE resampling 10 multiclass datasets.

The proposed method resampled dataset with SMOTE. Results indicate performance of the proposed method improved from 93% to 95%, RF improved from 93% to 94%, Naïve bayes improved from 87% to 86%, SVM improved from 70% to 76%, KNN improved from 85% to 86%, Bagging improved from 92% to 93%, JRipper improved from 82% to 85%, OneR improved from 69% to 70%, ZeroR remained at 50%, and C4.5 improved from 85% to 88%. Generally SMOTE resampling of datasets improved performance of all the algorithms. Results also indicate the proposed method outperformed Naïve bayes by 50%, SVM by 80%, KNN by 60%, JRipper by 70%, OneR by 100%, ZeroR by 100% and C4.5 by 70%. We also observe the proposed ensemble method outperformed ensemble bagging (Reptree) and ensemble Random forest algorithms.

Further observations review proposed method, ensemble random forest, ensemble bagging had better performance than other classification algorithms. Thus Ensemble technique produces more robust classifier that outperformed other algorithms.

Table 5: Statistical Paired T-test ROC Performance for Proposed Method, After SMOTE Resampling

Dataset	Proposed Method	RF	Naïve Bayes	SVM	KNN	Bagging	JRipper	OneR	ZeroR	C4.5
Cleverland	0.92	0.91	0.92	0.71*	0.81*	0.9	0.62*	0.7*	0.5*	0.87*
contraceptiveDataset	0.8	0.8	0.7*	0.72*	0.67*	0.81	0.71*	0.65*	0.5*	0.75*
dermatory	1	0.99	1	0.89*	0.94	0.98	0.95	0.61*	0.5*	0.96
ecolidataset	0.99	0.99	0.99	0.61*	0.96	0.99	0.95*	0.86*	0.5*	0.96*
myglassdata	0.96	0.94	0.76*	0.79*	0.82*	0.89*	0.82*	0.67*	0.5*	0.8*
newthyroid	0.99	0.99	0.99	0.92	0.97	0.98	0.94	0.88*	0.5*	0.96
RedWineQuality	0.94	0.93	0.82*	0.72*	0.82*	0.90*	0.82*	0.71*	0.5*	0.82*
vehicle	1	1	0.82*	0.54*	0.95*	0.99	0.95*	0.77*	0.5*	0.95*
yeastdataset	0.92	0.9	0.86*	0.73*	0.71*	0.88	0.81*	0.53*	0.5*	0.76*
Zoo	1	1	1	1	1	0.98	0.98	0.58*	0.5*	1
Average	0.95	0.94	0.88	0.76	0.86	0.93	0.85	0.7	0.5	0.88
	(x/ y/z)	(0/10/0)	(0/5/5)	(0/2/8)	(0/4/6)	(0/8/2)	(0/3/7)	(0/0/10)	(0/0/10)	(0/3/7)

10.3 Statistical Paired T-test between Proposed Method and Other Ensemble Algorithms

Table 6 presents the results. The results shows the proposed method outperformed individual ensemble algorithms used in the construction of the method. Further observations reveals the proposed method outperformed ensemble bagging (Reptree) by 20% and Ad_RF ensemble by 10%. Thus the proposed method outperformed other ensemble algorithms.

Table 6: Comparing Proposed Ensemble Method with other Ensemble Algorithms

Datasets	Proposed Method	AD_RF	RS_RF	RF	Bagging(Reptree)
Cleverland	0.92	0.92	0.92	0.91	0.90
Contraceptive	0.80	0.78	0.80	0.80	0.81
Dermatory	1	1	1	0.99	0.98
Ecoli	0.99	0.99	0.99	0.99	0.99
Glass	0.96	0.94	0.94	0.94	0.89*
Newthyroid	0.99	0.99	0.99	0.99	0.98
RedWine	0.93	0.92	0.93	0.93	0.90*
Vehicle	1	1	1	1	0.99
Yeast	0.9	0.89	0.9	0.9	0.88
Zoo	1	1	1	1	0.98
Average	0.95	0.94	0.95	0.94	0.93
Aggregation	(x/ y/z)	(0/9/1)	(0/10/0)	(0/10/0)	(0/8/2)

The proposed method was evaluated on the capability of outlier detection. Table 7 presents a summary of the

ROC metric performance for the minority classes using the 10 datasets. Results indicate the proposed method improved in outlier detection with all the datasets except the vehicle dataset.

.**Table 7:** ROC and F-Measure Outlier Metric Performance for 10 Multiclass Datasets using Proposed Method

Dataset	Outlier Class	F-Measure for Outlier Class	ROC measure for Outlier Class
Cleveland	Class 4	39.6% to 72.1%	80.6% to 94.2%.
Contraceptive	Long-term-method	39.6% to 72.1%	69.1% to 85.7%
Dermatology	Class 6(x6)	97.4% to 100%,	100% to 100%
Ecoli	clas 4 (imL)	Unknown to 99.2%	89.4% to 100%
Glass	Container	69.6% to 86.3%	97.2% to 99.3%
Redwine	Grade 8	16.4% to 94%	90.5% to 99.8%.
Vehicle	Van	75.7% to 73.3%	99.6% to 99.7%
Yeast	ERL	unknown to 66.9%	99.9% to 100%.
Zoo	Zoo5	97% to 97.9%	99.7% to 100%

11. Discussions

11.1 Feature Selection Methods and Outlier Detection in Data Mining

In this study, preprocessing was performed using an ensemble filter selection method and removal of point outliers. The ensemble filter selection method was constructed using Correlation, Gain-ratio, and Information-gain and ReliefF filter feature selection algorithms. A feature threshold was determined using Root Mean Square Error (RMSE) and random forest classifiers. The classifiers were evaluated using 10 fold stratified cross validation. The ensemble filter selection method improved the performance of classifiers. The study findings established that ensemble filter selection method produces a more robust filter method that improves on classification. The use of RMSE to determine the threshold makes the process statistically empirically testable and reliable.

Study by [5] used information-gain, gain-ratio, chi-square and ReliefF filter methods to create an ensemble filter selection method. Their study showed that combining feature selection methods improves the performance of classifiers by identifying the features that are weak as individual but strong as a group. Their findings affirmed that ensemble filter selection method result in an improved method that performs better than the individual selection methods used in the combination. However, this research study differs with the author's [5] study in the process of determining the threshold. The author used the output of the one-third split of ranked features to determine the feature to be selected. The method assumed that if a feature is voted by 3 methods, then it should be considered as a good feature. In this study we used RMSE to determine the threshold as discussed in the proposed method. This study findings is supported by [3] who demonstrated that feature selection methods improves knowledge of the classification process under consideration. Study by [4]

reaffirmed that ensemble learning can be realized through aggregating output of several models which results in a better results than the output of any individual model.. This study finding is in line with their conclusion.

Outlier detection approaches can be categorized into statistic-based, unsupervised, supervised, and semi-supervised. In this study, a model-based outlier detection was built using statistical IQR filter algorithm method. The study aimed at eliminating point outliers prior to building up the final model. The classifier built showed improved outlier detection after applying both the ensemble filter selection method and preprocessing through removal of point outliers. The results suggest removal of irrelevant or redundant features have an effect on reducing point-outliers leading to improved classifier performance. The study finding agrees with [37] who maintained that removing outlier improves outlier detection and performance of classification. Also this study findings reaffirms similar studies by [28] who reiterated that outliers influence the results of data mining techniques like clustering, classification and association.

11.2 Multiclass Classification and Outliers Detection Methods in Data Mining

The study sought to address the issue of imbalance problem associated with multiclass. Existing common classification learning algorithms are biased towards the prediction of the majority classes. Sampling techniques such as under-sampling, oversampling and SMOTE can be applied to balance datasets for .Minority classes in a multiclass problem may be described as rare classes or rare events or outliers [6]. In this study, multiclass datasets were individually resampled using synthetic minority over sampling technique (SMOTE) to generate synthetic instances from the minority classes. The numbers of artificial samples were generated to the level that measured at least 50% match with the majority classes. The model built showed improvements in the classification and outlier detection performance after resampling the datasets. This study avoided the use of over-sampling which tends to increase the number of the minority instances that can result to over-fitting because of the duplication of data. Also the study did not advocate the use of under-sampling which reduce the number of the majority instances but may result to a loss of information of the majority and hence decrease the performance of classification. [6] established that SMOTE improves on data representation that improves on performance of classifiers. They evaluated their model using ROC metrics performance of C4.5, Ripper and Naïve bayes classifiers. This study findings agrees with the conclusion of the author. Furthermore [19] demonstrated that the SMOTE improves on the generalization performance of classifiers which is reaffirmed by results of this study.

In this study, each of the 10 multiclass datasets were transformed to binary classes using OnevsOne decomposition technique utilizing pairwise coupling. The model built performed well in classification and outlier detection. This study finding is in line with [1] who supported the techniques of extending binary classification problems to handle the multiclass. In their findings, they affirmed that a multiclass problem can be decomposed into several binary classes and the resulting classifier performs well. The technique used in this study is in conformity with study by [26] who proclaimed that researchers use multiclass decomposition technique of one against one, and one against all to address the issue of multiclass classification.

11.3 Use of Ensemble Learning Methods

In this study, a heterogeneous ensemble model was built using ensemble Adaboost and ensemble Random subspace algorithms using random forest as the base classifiers for each of the algorithms. According to [24] combining ensemble learning with resampling techniques produce better classifiers that outperform other individual classic algorithms. In this study, adaboost was considered in the model due to its boosting capabilities. According to [10] boosting works by repeatedly running a given weak learning algorithm on various distributions over the training data, and then combining the classifiers produced by the weak learner into a single composite classifier.

Random subspace algorithm was considered in this study since it benefits from bootstrapping and aggregation focusing on feature space. The original training set is modified but unlike bagging that bootstraps training samples, this modification is performed in the feature space by projecting the original feature space into different and randomly chosen feature subspaces. According to [20], RMS can effectively handle datasets with redundant features and it has been known to overcome the problems of instability and over-fitting.

Random forest algorithm was considered in this study since it is the most commonly used ensemble classification algorithm and computes the importance of each feature in the classification process. The algorithm utilizes the combination of tree predictors to produce tree classifiers. The algorithm use bagging and random selection of features to split at each node. The study aimed at boosting the performance of classifier built using random forest as the base classifier.

11.4 Performance of the Proposed Method

Statistical Paired-T test was used to compare the performance of the proposed method. Comparisons were performed with the individual classifiers used in the method and also with other existing algorithms. Results showed Ad_RF and Random forest statistically lost once to the proposed method but the proposed method had zero significantly error rate to RS_RF algorithm. Further observation showed the ensemble random forest, ensemble bagging had better performance than other classification algorithms though they were outperformed by the proposed method. This study support conclusion of [18] who reaffirmed that random subspace method (RSM) is a good ensemble learning algorithms and is widely used in pattern classification applications. The authors reiterated that RSM has the advantages of small error rate and improved noise insensitivity due to ensemble construction of the base-learners. The use of ensemble method in this study is in line with assertion by [17] who claimed that many ensemble methods, seek to promote diversity among the models they combine. Ensembles tend to yield better results when there is a significant diversity among the models.

The technique used in this study is slightly different from the one proposed by [34] who constructed heterogeneous ensemble classifier using combination of bagging, boosting and random subspace ensemble algorithms. The author used model tree inducer M5rules in one set of experiment and sequential minimal optimization (SMO) algorithm in another set of experiment whereas in this study we used random forest algorithm as the base learner. Another difference arises in the handling of the multiclass. In this study we applied SMOTE resampling and decomposed the datasets using OnevsOne technique. The Authors did not balance the datasets nor attempted to remove point outliers that could be associated with some attributes of some

of the datasets. Also the author only carried a comparative study with the individual classifiers used in the construction of the ensemble. In this study, we compared the performance of the proposed heterogeneous ensemble method with 10 other well-known classification algorithms and performed preprocessing using an ensemble filter method that we created.

12. Conclusion

This study proposed the use of ensemble filter selection method to improve on the performance of classifiers and rebalancing datasets using SMOTE. The ensemble filter method provides an improved reliable effective way of preprocessing datasets. Since multiclass problems are well handled through binarization, this research proposed use of decomposition technique rather than the technique of merging classes. Also we propose removal of point-outliers associated with features prior to the building of classification and outlier classifiers. This research proposed development of heterogeneous ensemble model using boosting and bagging techniques.

In this study, model-based outlier detection method was built using statistical IQR filter algorithm method. The results suggest removing irrelevant or redundant features have an effect of reducing point-outliers leading to improved classifier performance. The study demonstrates that presence of outlier degrades and influences the performance of classifiers. The study demonstrates resampling dataset using SMOTE improves on rare class detection. The study confirmed ensemble techniques produce better outlier detection and classification performance. The study findings reaffirmed that ensemble learning can be realized through aggregating output of several models which results in a better results than the output of one individual model.

13. Recommendation

In this study, an ensemble filter selection method was developed and used in the development of the proposed model using four filter selection algorithms. There is need to explore the effect of creating an ensemble selection method using combination of filter and wrapper selection methods. Future study should investigate the effect of partial removal of point-outliers from datasets prior to building up of classifiers. In this study, skewed datasets were resampled using SMOTE algorithm to the extent that the resulting minority class distribution measured at least 50% match in comparison with the majority class. Further study could be done to ascertain an ideal percentage of resampling rather than choosing an arbitrary distribution measure. In this study the model built utilized the capabilities of boosting and bagging ensemble learning algorithms. We recommend further study on combining the proposed method with deep learning algorithms.

References

[1] M. Elkano, M. Galar, J. Sanz , G. Lucca & H. Bustince. IVOVO: “A new interval-valued one-vs-one approach for multi-class classification problems”. In2017 Joint 17th World Congress of International Fuzzy Systems Association and 9th International Conference on Soft Computing and Intelligent Systems (IFSA-SCIS) 2017 Jun 27 (pp. 1-6). IEEE.

[2] K. Malik, & H.G.S.K. Sadawarti, “ Comparative analysis of outlier detection techniques”. International

Journal of Computer Applications. 2014 Jul;97(8):12-21.

[3] S., Khalid , T. Khalil, & S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning". In2014 Science and Information Conference 2014 Aug 27 (pp. 372-378). IEEE.

[4] B., Seijo-Pardo, I. Porto-Díaz, V. Bolón-Canedo, &A. Alonso-Betanzos, " Ensemble feature selection: homogeneous and heterogeneous approaches". Knowledge-Based Systems. 2017 Feb 15;118:124-39.

[5] O. Osanaiye, H. Cai, KK. Choo, A. Dehghantanha, Z. Xu, & M. Dlodlo, "Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing". EURASIP Journal on Wireless Communications and Networking. 2016 Dec;2016(1):130.

[6] N.V. Chawla, K.W. Bowyer, L.O. Hall, & W.P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique". Journal of artificial intelligence research. 2002 Jun 1;16:321-57.

[7] MY. Kiang, "A comparative assessment of classification methods". Decision support systems. 2003 Jul 1;35(4):441-54.

[8] TN. Phyu,. "Survey of classification techniques in data mining". In Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Mar 18 (Vol. 1, pp. 18-20).

[9] S. Chen, G. Guo, &L. Chen, "A new over-sampling method based on cluster ensembles". In 2010 IEEE 24th International Conference on Advanced Information Networking and Applications Workshops 2010 Apr 20 (pp. 599-604). IEEE.

[10] MK. Khan, & A. Umer, "An Experimental Evaluation of Ensemble Methods for Pattern Classification". In 2011 Third International Conference on Computational Intelligence, Communication Systems and Networks 2011 Jul 26 (pp. 6-10). IEEE.

[11] S. Wang, & X. Yao, "Multiclass imbalance problems: Analysis and potential solutions". IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics). 2012 Aug;42(4):1119-30.

[12] HY. Lin, "Efficient classifiers for multi-class classification problems". Decision Support Systems. 2012 Jun 1;53(3):473-81.

[13] R. Longadge, &S. Dongre, "Class imbalance problem in data mining review". arXiv preprint arXiv:1305.1707. 2013 May 8.

[14] MA. Bagheri, Q. Gao, &S. Escalera, "A framework towards the unification of ensemble classification methods". In2013 12th International Conference on Machine Learning and Applications 2013 Dec 4 (Vol. 2, pp. 351-355). IEEE.

[15] Y. Ming-hai, & W. Na, "Research on the ensemble learning classification algorithm based on the novel feature selection method". In Proceedings of 2013 IEEE International Conference on Vehicular Electronics and Safety 2013 Jul 28 (pp. 263-267). IEEE.

[16] U. Turhal, S. Babur, C. Avci, &A, Akbaş, "Performance improvement for diagnosis of colon cancer by using ensemble classification methods". In 2013 The International Conference on Technological Advances in Electrical, Electronics and Computer Engineering (TAEECE) 2013 May 9 (pp. 271-275). IEEE.

[17] K. Fawagreh, MM. Gaber, &E. Elyan, "Random forests: from early developments to recent advancements. Systems Science & Control Engineering": An Open Access Journal. 2014 Dec

1;2(1):602-9.

[18] A. Mert, N. Kılıç, &E. Bilgili, "Random subspace method with class separability weighting". Expert Systems. 2016 Jun;33(3):275-85.

[19] C. Gong, &L. Gu, "A novel SMOTE-based classification approach to online data imbalance problem". Mathematical Problems in Engineering. 2016;2016.

[20] I. Barandiaran, "The random subspace method for constructing decision forests". IEEE Trans. Pattern Anal. Mach. Intell. 1998 Aug;20(8).

[21] Q. Zhang, editor. Visual Analytics and Interactive Technologies: Data, Text and Web Mining Applications: Data, Text and Web Mining Applications. IGI Global; 2010 Oct 31.

[22] K. Singh , & S. Upadhyaya, "Outlier detection: applications and techniques". International Journal of Computer Science Issues (IJCSI). 2012;9(1):307.

[23] J. Zhang, "Advancements of outlier detection: A survey". ICST Transactions on Scalable Information Systems. 2013 Feb 4;13(1):1-26.

[24] W. Feng, W. Huang, & J, Ren, "Class imbalance ensemble learning based on the margin theory". Applied Sciences. 2018 May;8(5):815.

[25] K. Chomboon, K. Kerdprasop, &N. Kerdprasop, "Rare class discovery techniques for highly imbalance data". In Proc. International multi conference of engineers and computer scientists 2013 (Vol. 1).

[26] N. Mehra, & S. Gupta , "Survey on multiclass classification methods".

[27] SS. Sreevidya, "Detection of outliers in data stream using clustering method". Int J Sci Eng Technol Res. 2015;4(3):559-63.

[28] L. Sunitha , M. BalRaju, J. Sasikiran, & EV. Ramana," Automatic outlier identification in data mining using IQR in real-time data". International Journal of Advanced Research in Computer and Communication Engineering. 2014;3(6):7255-7.

[29] M. Cárdenas-Montes, MA. Vega-Rodríguez, JJ. Rodríguez-Vázquez, &A. Gómez-Iglesias," A comparison exercise on parallel evaluation of rosenbrock function". In Proceedings of the Companion Publication of the 2015 Annual Conference on Genetic and Evolutionary Computation 2015 Jul 11 (pp. 1361-1362). ACM.

[30] B. Liu , Y. Xiao , SY. Philip, Z. Hao, &L. Cao , " An efficient approach for outlier detection with imperfect data labels". IEEE transactions on knowledge and data engineering. 2014 Jul;26(7):1602-16.

[31] SS. Rakhe, &AS. Vaidya,"A Survey on Different Unsupervised Techniques to Detect Outliers". International Research Journal of Engineering and Technology (IRJET) Volume. 2015;2.

[32] A. Christy, GM. Gandhi, &S. Vaithyasubramanian, "Cluster based outlier detection algorithm for healthcare data". Procedia Computer Science. 2015 Jan 1;50:209-15.

[33] M. Radovanović, A. Nanopoulos, &M. Ivanović, " Reverse nearest neighbors in unsupervised distance-based outlier detection". IEEE transactions on knowledge and data engineering. 2015 May 1;27(5):1369-82.

[34] S. Kotsiantis, &D.Kanellopoulos, "Combining bagging, boosting and random subspace ensembles for regression problems". International Journal of Innovative Computing, Information and Control. 2012 Jun 1;8(6):3953-61.

[35] K. Yan, X. You, X, Ji, G, Yin &F. Yang, “A hybrid outlier detection method for health care big data” . In 2016 IEEE International Conferences on Big Data and Cloud Computing (BDCloud), Social Computing and Networking (SocialCom), Sustainable Computing and Communications (SustainCom)(BDCloud-SocialCom-SustainCom) 2016 Oct 8 (pp. 157-162). IEEE.

[36] R. Bansal, N. Gaur, & SN. Singh,” Outlier detection: applications and techniques in data mining”. In 2016 6th International Conference-Cloud System and Big Data Engineering (Confluence) 2016 Jan 14 (pp. 373-377). IEEE.

[37] D. Singh , EJ. Leavline, “ Model-Based Outlier Detection System with Statistical Preprocessing”. Journal of Modern Applied Statistical Methods. 2016;15(1):39.

[38] D. Dua, E. Karra Taniskidou ,” UCI Machine Learning Repository” [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California. School of Information and Computer Science. 2017.