Journal of Information Engineering and Applications www.iiste.org
ISSN 2224-5782 (print) ISSN 2225-0506 (online) DOI: 10.7176/JIEA
Vol.9, No.2, 2019

An Ensemble Model for Multiclass Classification and Outlier Detection Method in Data Mining

Dalton Ndirangu[1*] Prof. Waweru Mwangi[2] Dr. Lawrence Nderu[2]
1.United States International University-Africa, P.O. Box 14634 00800, Nairobi, Kenya
2.Jomo Kenyatta University of Agriculture and Technology, P.O. Box 62,000 – 00200 Nairobi, Kenya

Abstract
Real life world datasets exhibit a multiclass classification structure characterized by imbalance classes. Minority classes are treated as outliers' classes. The study used cross-industry process for data mining methodology. A heterogeneous multiclass ensemble was developed by combining several strategies and ensemble techniques. The datasets used were drawn from UCI machine learning repository. Experiments for validating the model were conducted and represented in form of tables and figures. An ensemble filter selection method was developed and used for preprocessing datasets. Point-outliers were filtered using Inter quartile range filter algorithm. Datasets were resampled using Synthetic minority oversampling technique (SMOTE) algorithm. Multiclass datasets were transformed to binary classes using OnevsOne decomposing technique. An Ensemble model was developed using adaboost and random subspace algorithms utilizing random forest as the base classifier. The classifiers built were combined using voting methodology. The model was validated with classification and outlier metric performance measures such as Recall, Precision, F-measure and AUCROC values. The classifiers were 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 well-known existing classification and outlier detection algorithms such as Naïve bayes, KNN, Bagging, JRipper, Decision trees, RandomTree and Random forest. The study findings established ensemble techniques, resampling datasets and decomposing multiclass results in an improved detection of minority outlier (rare) classes.
Keywords: Multiclass, Outlier, Ensemble, Model, Classification
DOI: 10.7176/JIEA/9-2-04
Publication date: April 30th 2019

1. INTRODUCTION

Outlier detection has continued to be an active research area in the field of data mining due to its challenges and wide application. (Wang & Huang, 2017) affirms that outlier detection is one the important research area of data mining, which plays vital role in data cleansing and detection of rare events or abnormal incidents. From a statistical perspective, outliers are described as those observations that are significantly different from the majority. Depending on the specific applications, outliers are also referred to as anomalies, discordant observations, exceptions, faults, defects, aberrations, noise, errors, damage, surprise, novelty, peculiarities or contaminants(Chandola, et. al, 2007).

With the emerging technologies such as cloud computing, internet of things and social networks, outlier detection has continued to obtain wide applications. Outlier detection provide useful, sufficient and meaningful knowledge in data preprocessing, equipment fault diagnosis, credit fraud detection, traffic incident detection, network intrusion and others (Chandarana, 2015) According to (Bansal, Gaur, & Singh, 2016), outlier detection is one of the major issues in data mining. Outliers may appear as a deviation from the rest of the objects as if it was generated from a different mechanism (Hawkins, 1980) or may be represented as rare events or minority class in a classification problem (Seiffert, 2007).

Real world problem is characterized by presence of multiclass problem. The latter is more often than not composed of imbalance dataset representation. The rare classes' form the class of interest since the existing classification algorithms were designed with bias towards prediction of majority classes. Thus several strategies and techniques are required when solving the problem of multiclass. According to (Elkano, Galar, Sanz, Lucca, & Bustince, 2017) decomposition strategies have been demonstrated to be a successful methodology for multiclass classification problems.

Authors (Lin & Yan, 2015) assert that classification and prediction through data mining can grasp the basic trend of the development of the unknown data. Since outliers can manifest themselves as rare events in a multiclass classification problem, it is practical to combine the study of multiclass classification and outlier detection. Thus a novel prediction method should be developed that improves on the prediction of the minority classes and safeguard the integrity performance of the majority classes.

The next section 2 provides related work while the proposed method is presented in section 3. Section 4 provide experiments and analysis while section 5 concludes the paper.

2.0 RELATED WORKS

Data mining techniques have been applied and integrated on several fields such as machine learning, statistics, artificial intelligence, and database systems, for analysis of large volumes of data (Allahyari, Trippe, & Gutierrez, 2017). According to (Han, 2015), predictive activities of data mining include classification and regression. The primary objective of supervised learning is to build a concise model of the distribution of class labels in terms of predictor features (Kotsiantis & Kanellopoulos, 2012).

Supervised outlier detection techniques assume the availability of a training data set which has labeled instances for normal as well as outlier class in a classification problem (Gogoi, Bhattacharyya, Borah, & Kalita, 2011). Typical practical approach in such case is to build predictive models for both normal and outlier classes. Any unseen data example is compared against the two models to determine the appropriate class it belongs to. Supervised outlier detection techniques have an explicit notion of the normal and outlier behavior and hence accurate classifiers can be built (Chandola et al., 2007).

(Zhang, 2013) proclaimed outlier detection is an important research problem in data mining that attempts to discover 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. Different models or algorithm may detect and output outliers differently (Rana, Pahuja, & Gautam, 2014) and hence the need to develop an effective outlier detection method.

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 (J. Wang, Zhou, Yi, & Kong, 2014).

According to (Nikulin & Mclachlan, 2009), significant progress in classifier metric performance is achieved through the advanced preprocessing and feature selection techniques. Authors concluded that use of several different models results in an improved classifier. The models could be improved through ensemble technique. An ensemble method is considered as meta-algorithm that results from combination of several machine learning algorithms and techniques that aims at reducing data variance, algorithm bias and improves prediction. 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 (Oza, 2000).

It has been noted most of the ensemble methods use a single base learning algorithm to produce homogeneous base learners although some methods that use learners of different types leading to heterogeneous ensembles. Author (Breiman, 2001) reaffirmed that 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. Recent studies have shown that combining feature selection methods through ensemble technique improves performance of classifiers by identifying features that are weak as an individual but strong as a group (Osanaiye et al., 2016).

Data sampling is a useful procedure when the data to be analyzed is of imbalanced class distribution where the samples from majority class outnumber samples from minority class (Feng, Huang, & Ren, 2018). Due to the inherent complex characteristics of imbalanced datasets, learning from such data requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently into information and knowledge representation.

Multiclass problem issues can be addressed through data resampling, decomposition and improving on the learning algorithms. Resampling techniques such as under- sampling, over-sampling, and synthetic minority oversampling technique (SMOTE.) are widely used to rebalance datasets. However more research is needed for the ever challenging emerging multiclass problem in real life applications(Krawczyk, 2016).

3.0 CONCEPTUAL FRAMEWORKS

The conceptual framework in this study was based on input-process-output model. Figure 4 shows the framework. The input is made up of multiclass datasets. The process include preprocessing, multiclass transformation, building of classifiers, and combing classifiers. Output include predicted classification and outlier (minority classes) performance.

The dataset is filtered to remove redundant or irrelevant features. An ensemble filter selection method is developed using combination of several filter selection algorithms. The dataset is split into training and testing samples. The split can be 70% for training and 30% for testing. Alternatively Stratified 10 fold cross validation is used. The initial dataset is partitioned into 10 subsets with an approximately equal number of samples in each subset. Each subset is used as the test partition, while the remaining subsets is considered for training dataset.

The filtered dataset is preprocessed through feature reduction and global outlier removal. Some of the resampling technique that can be applied include, under-sampling, over-sampling, generating artificial samples using SMOTE, etc. Removal of the outliers can be done through use of model-based outlier detection using statistical IQR method.

The preprocessed dataset is transformed to binary using any one of the several decomposition techniques such

as OVA, OVO, and ECOC. Ensemble classifiers are built using the selected data set features and learning classification algorithms. Single or several learning algorithms are modeled with the same or different base (weaker) learners. The resulting classifiers are combined using voting methodology. Some of the voting methodology used include majority voting, minimum probability, maximum probability, median, average of probabilities and product of probabilities.

Classification and outlier performance is achieved through predictive analysis using ROC measure, detection rate (true positive), precision, recall and F-measure.

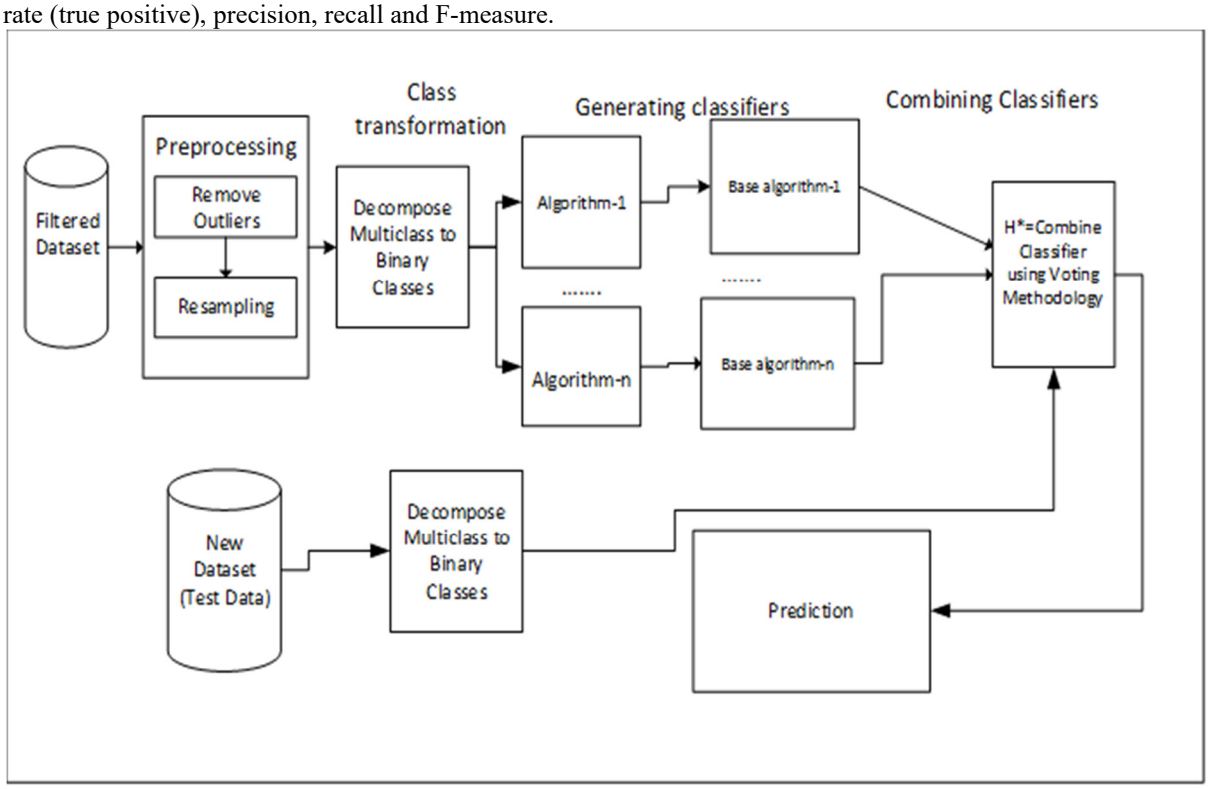

Figure 1: Conceptual Framework of the Study

4.0. PROPOSED METHOD

From the conceptual framework described in section 3, we proposed development of an ensemble multiclass classification and outlier detection method for data mining. The method used several strategies and ensemble techniques. The method had six phases namely, development of an ensemble filter selection method phase, data preprocessing phase, dataset resampling and point-outlier filtering phase, multiclass transformation phase, ensemble model building, testing and validation phase. To demonstrate the significant of our method, we used multiclass datasets. Initial exploration of the datasets involved plotting histogram to ascertain presence of point outliers. The phases are described as shown in the next subsections:

4.1 Developing an Ensemble Filter Feature Selection Method

Phase 1 involved development of an ensemble filter feature selection method that was to be used as part of preprocessing. The process used Correlation, Information-gain, ReliefF, and Gain-ratio algorithms. The algorithms were considered due to the fact that Correlation algorithm aims at establishing a feature list that has lesser feature-feature correlation with each other and higher feature-class correlation. Information gain algorithm is known to evaluate the worth of an attribute by measuring the information gain with respect to the class. Gain ratio on the other hand evaluates the worth of an attribute by measuring the gain ratio with respect to the class. ReliefF algorithm can easily deal with multiclass problems and is also more robust and capable of dealing with incomplete and noisy data.

Figure 2 provides the pseudo-code for developing the ensemble filter selection method. Each of the four algorithms were selected and used to individually rank the features of the datasets. This resulted in generation of four ranked feature lists. The four lists were then sorted and merged using aggregation or majority voting techniques. Random forest classifiers and Root Mean Square Error (RMSE) were used to determine the relevant optimal features in the merged list. The process started by building classifier using the top-ranked feature in the merged list and the resulting RMSE value observed and recorded. The process was repeated iteratively by incorporating the next top-most feature. As long as a feature had significant contribution to the performance of

classifiers, the RMSE predictive value was expected to continue decreasing as more bottom ranked features were incorporated. When a feature with less contribution to the performance of classifier was incorporated, the resulting classifier was expected to have a higher RMSE value compared with the previous immediate RMSE value. Thus the threshold was set to this level where the classifiers started to deteriorate in terms of performance. The final expected feature sub-list included the features starting from the top-ranked feature up to and including the feature that resulted in the generation of the least RMSE value.

Algorithm: Creating Ensemble Filter Selection Method

Input: **M** is number of filter selection methods, **F** is the number of Features in Dataset **D**, **L** is Lists of ranked features, **MList** is a merged list of features,

Output: **R** is store for RMSE, **S** is subset of optimal ensemble list of features

Procedure:

Assume **L**, **R** and **S** is empty,

1 Apply filter selection method f_i, i=1 to **M**, to rank all the Features **F** in Dataset **D** to produce;
- 1.1 $L_i=\{fj_1,\ldots,fjn\}$ where fj equals ranked feature where n is less or equal to **F** and i=1 to **M**
- 1.2 Store ranked features to $L=\{L_1,\ldots,Lm\}$ where $m \in \mathbf{M}$
- 1.3 Merge lists **L** of ranked features to produce **MList**
- 1.4 Using **MList** and Starting from the top feature Do:
  - 1.4.1 Select a feature from **MList** and build a random forest classifier (model)
  - 1.4.2 Compute current RMSE value of the model, say Ri where i≤F
  - 1.4.3 Compare current RMSE value with the previous RMSE Value, If greater, Stop else
  - 1.4.4 Store current RMSE value to **R**=**R** U{Ri}
  - 1.4.5 Store selected feature to **S**= **S**U {Si}
  - 1.4.6 Select the Next feature in the **MList**
- 1.5 Output **R** and **S**

Figure 2: Pseudo-code for Creating Ensemble Filter Selection Method

4.2 Preprocessing Datasets

Phase 2 involved further data preprocessing using the developed ensemble filter selection method and filtering point outliers using interquartile range (IQR) algorithm. The outliers were identified from statistical tail ends as follows:

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

Where: Q_1 = 25% quartile, Q_3 = 75% quartile and IQR = difference between Q_1 and Q_3

4.3 Resampling Datasets

Phase 3 involved resampling the datasets. SMOTE was used to generate artificial samples of the rare classes. The numbers of artificial samples were generated to the level that they measured at least 50% compared with the majority classes.

4.4 Decomposing Multiclass

Phase 4 involved transforming multiclass problem to binary problem. The proposed method used One-verses-One technique utilizing pairwise coupling to speed up the decomposition process.

4.5 Building Ensemble Model

Phase 5 involved building a heterogeneous ensemble model. Two ensemble classifiers AD_RF and RS_RF were built using Adaboost algorithm and Random Subspace algorithm respectively each utilizing random forest algorithm as their base classifier. The two ensemble classifiers were combined using voting technique utilizing average of probabilities combination rule. Each individual classifier (AD_RF, RS_RF) generated their hypothesis h1, and h2 respectively. For each output class, a posteriori probabilities was generated by individual classifier AD_RF and RS_RF. Thereafter, the class represented by the maximum average value of a posteriori probabilities was designated to be the voting hypothesis (h*) for the final decision outcome. Figure 3 provide the proposed ensemble model.

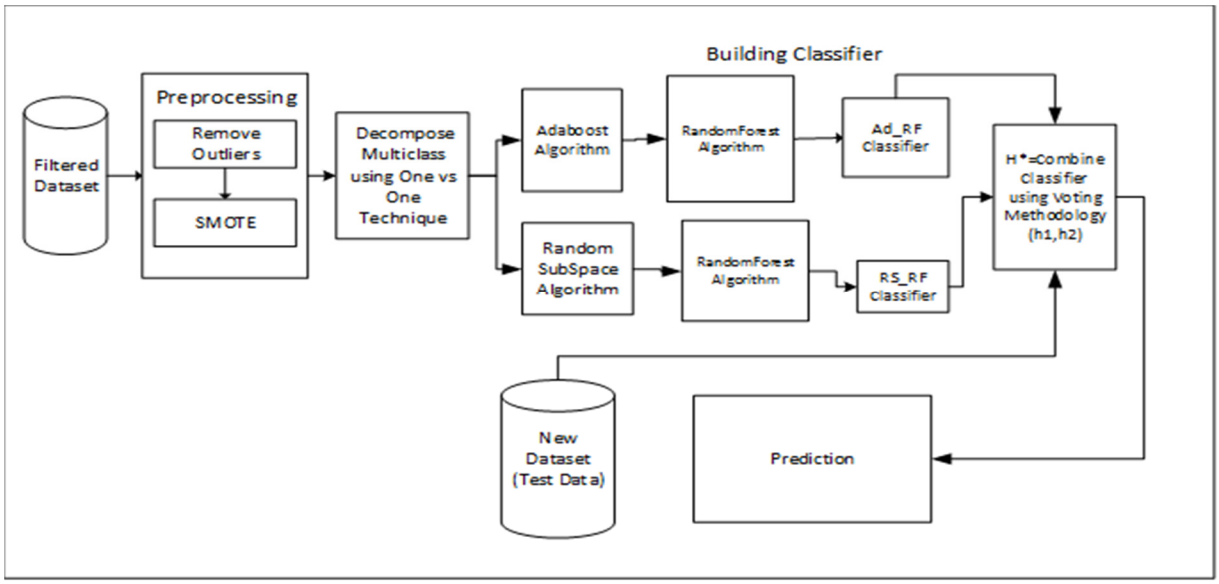

Figure 3: The Proposed Ensemble Model

4.6 Testing and Validating the Model

Phase 6 involved testing the metric performance of the proposed method. Stratified 10 fold cross validation was used to valid the performance of the model. The initial dataset was first partitioned into 10 subsets with an approximately equal number of records in each subset. Each subset was used as the test partition, while the remaining subsets were combined to perform the role of the training partition. A paired T-test was employed for testing the difference in performances of the proposed model and the commonly used classification algorithms. The paired T – test applied 95% statistical confidence interval. Several tests were performed including testing the merit for each module in the method and the comparison metric performance of the method with the other well-known classification algorithms. Receiver Operating Characteristic (ROC) values was used to measure the performance of the classifiers. Other metrics performance measures such as True Positive, Precision, Recall and F-measure were used to evaluate the performance of model.

5.0 EXPERIMENTS AND ANALYSIS

5.1 Dataset Description

UCI (Dua, D. and Karra Taniskidou, 2017) database is of high-quality, real-world, and well understood machine learning datasets. Since the study focused on multiclass and outlier detection, all the datasets drawn from UCI were multiclass. The datasets were a mixture of low and high dimensional data with varying number of instances, and classes. Majority of the datasets did not have missing values. All the datasets had imbalance classes. Table 4.1 shows summary description of the datasets after applying the proposed ensemble filter feature selection method.

Table 4.1: Summary Description of Datasets and Selected Features

	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

4.2 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. Table 4.2 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 proposed method and other existing algorithms.

Table 4.2: 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

4.3 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 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 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" represented the winning situation of that particular algorithm as compared with the proposed algorithm while "*" indicated that the proposed ensemble algorithm was statistically better than the compared algorithm. **Plain text** signified that there was no difference in performance indicating a draw. Aggregated results are represented in terms of x, y, and z where "x" represented number of aggregated losses and "z" represents aggregated number of wins and "y" represents aggregated number of draws for the proposed method.

Table 4.3 shows result of experiment after SMOTE resampling 10 multiclass datasets while Figure 4 shows effect of SMOTE on classifiers. The proposed method advocated use of resampling dataset with SMOTE. Performance was measured using ROC values. Results indicate performance of the proposed method improved attained 95%, RF attained 94%, Naïve bayes registered 86%, SVM had 76%, KNN attained 86%, Bagging registered 93%, JRipper had 85%, OneR attained 70%, ZeroR had 50%, and C4.5 attained 88%.

Generally SMOTE resampling of datasets improved performance of all the algorithms as shown in Figure 4.2. 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 outperforms other algorithms.

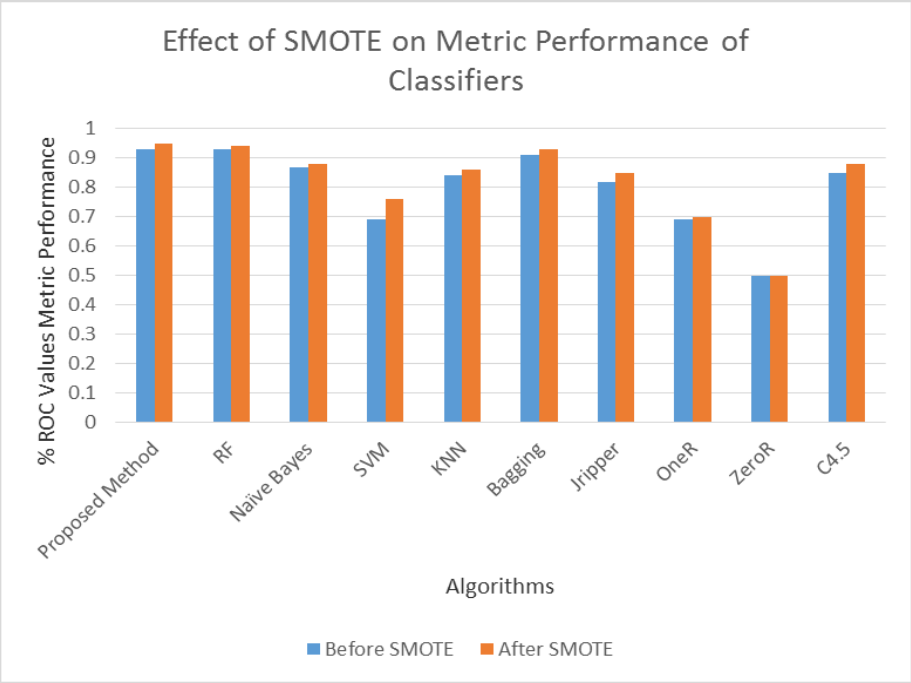

Figure 4: Effect of SMOTE on Performance of Classifiers

Table 4.3: Statistical Paired T-test ROC Performance for Proposed Method, After SMOTE Resampling

Dataset	Proposed Method	RF	Naïve Bayes	SVM	KNN	Bagging	JRipper	OneF	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*
vehide	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)

4.4 Statistical T-test between Proposed Method and Other Ensemble Algorithms

Table 4.4 presents the results. The results shows the proposed method outperformed individual ensemble algorithms used in the construction of the ensemble method. Further observations reveals the proposed method outperformed ensemble bagging (Reptree) by 20% and Ad_RF ensemble by 10%.

Table 4.4: 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)

5. CONCLUSION

A heterogeneous ensemble model was developed using adaboost and random subspace algorithms both utilizing random forest as the base classifiers. The model incorporated effective ways of preprocessing datasets through ensemble filter method and resampling datasets using SMOTE algorithms. To further increase the predictability of minority outlier classes, multiclass datasets were decomposed to binary classes using OnevsOne technique. Since point-outliers degrade performance of classifiers, the model built had a preprocessing mechanism using IQR outlier filter algorithm. The performance of the proposed model was compared with other existing well known classification algorithms using metric performance measures of Recall, Precision, and ROC values. The model built outperformed most of the existing classification and outlier detection algorithms. We conclude that ensemble technique through feature selection and combining algorithms produce a more robust classifier.

References

Allahyari, M., Trippe, E. D., & Gutierrez, J. B. (2017). A Brief Survey of Text Mining : Classification , Clustering and Extraction Techniques.

Bansal, R., Gaur, N., & Singh, S. N. (2016). Outlier Detection: Applications and techniques in Data Mining. *2016 6th International Conference - Cloud System and Big Data Engineering (Confluence)*, 373–377. https://doi.org/10.1109/CONFLUENCE.2016.7508146

Breiman, L. E. O. (2001). Random Forests, 5–32.

Chandarana, D. R. (2015). A Survey for Different Approaches of Outlier Detection in Data Mining.

Chandola, V., Banerjee, A., & Kumar, V. (2007). Outlier detection: A survey. *ACM Computing Surveys*, *14*, 15.Dua, D. and Karra Taniskidou, E. (2017). UCI (University of California Irvine) Machine Learning Repository.

Elkano, M., Galar, M., Sanz, J., Lucca, G., & Bustince, H. (2017). IVOVO : A new Interval-Valued One-Vs-One approach for multi-class classification problems.

Feng, W., Huang, W., & Ren, J. (2018). Class Imbalance Ensemble Learning Based on the Margin Theory. *Applied Sciences*, *8*(5), 815. https://doi.org/10.3390/app8050815

Gogoi, P., Bhattacharyya, D. K., Borah, B., & Kalita, J. K. (2011). A Survey of Outlier Detection Methods in Network Anomaly Identification, *54*(4). https://doi.org/10.1093/comjnl/bxr026

Han, J. (2015). Data Mining : Concepts and Techniques.

Kotsiantis, S., & Kanellopoulos, D. (2012). Combining bagging, boosting and random subspace ensembles for regression problems. *International Journal of Innovative Computing, Information and Control*, *8*(6), 3953–3961.

Krawczyk, B. (2016). Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, *5*(4), 221–232. https://doi.org/10.1007/s13748-016-0094-0

Lin, C., & Yan, F. (2015). The study on classification and prediction for data mining. *2015 Seventh International Conference on Measuring Technology and Mechatronics Automation*, 1305–1309. https://doi.org/10.1109/ICMTMA.2015.318

Nikulin, V., & Mclachlan, G. J. (2009). Classification of Imbalanced Marketing Data with Balanced Random Sets, 89–100.

Osanaiye, O., Cai, H., Choo, K.-K. R., Dehghantanha, A., Xu, Z., & Dlodlo, M. (2016). Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing. *EURASIP Journal on Wireless Communications and Networking*, *2016*(1), 130. https://doi.org/10.1186/s13638-016-0623-3

Oza, N. C. (2000). Online Ensemble Learning, 2000.

Rana, P., Pahuja, D., & Gautam, R. (2014). A Critical Review on Outlier Detection Techniques, *3*(12), 2394–2403.

Seiffert, C. (2007). Mining Data with Rare Events : A Case Study, 132–139. https://doi.org/10.1109/ICTAI.2007.71

Wang, J., Zhou, S., Yi, Y., & Kong, J. (2014). An Improved Feature Selection Based on Effective Range for Classification, *2014*.

Wang, Z., & Huang, X. (2017). An Outlier Detection Algorithm Based on the Degree of Sharpness and Its Applications on Traffic Big Data Preprocessing, 478–482.

Zhang, J. (2013). Advancements of Outlier Detection: A Survey. *ICST Transactions on Scalable Information Systems*, *13*(1), e2. https://doi.org/10.4108/trans.sis.2013.01-03.e2