10.31341/jios.47.1.4 UDC 004.738.5:316.77
Open Access Original Scientific Paper

A Modified Boosted Ensemble Classifier on Location Based Social Networking

Lakshmi Shree K *lakshmishreek@vemanait.edu.in*
Research Scholar, Department of Information Science and Engineering
BMS College of Engineering, Bangalore, India

Ranganath Ashok Kumar *ashokkumar.ise@bmsce.ac.in*
Professor, Department of Information Science and Engineering
BMS College of Engineering, Bangalore, India

Abstract

One of the research issues that researchers are interested in is unbalanced data classification techniques. Boosting approaches like Wang's Boosting and Modified Boosted SVM (MBSVM) have been demonstrated to be more effective for unbalanced data. Our proposal The Modified Boosted Random Forest (MBRF) classifier is a Random Forest classifier that uses the Boosting approach. The main motivation of the study is to analyze sentiment of geotagged tweets understanding the state of mind of people at FIFA and Olympics datasets. Tree based model Random Forest algorithm using boosting approach classifies the tweets to build a recommendation system with an idea of providing commercial suggestions to participants, recommending local places to visit or perform activities. MBRF employs various strategies: i) a distance-based weight-update method based on K-Medoids ii) a sign-based classifier elimination technique. We have equally partitioned the datasets as 70% of data allocated for training and the remaining 30% data as test data. Our imbalanced data ratio measured 3.1666 and 4.6 for FIFA and Olympics datasets. We looked at accuracy, precision, recall and ROC curves for each event. The average AUC achieved by MBRF on FIFA dataset is 0.96 and Olympics is 0.97. A comparison of MBRF and Decision tree model using 'Entropy' proved MBRF better.

Keywords: Social Network Services (SNS), Location Based Social Media (LBSM), Random Forest, Imbalanced data learning

1. Introduction

Users quickly adopted the global popularity of Social Network Services (SNS) such as Foursquare and Yelp, and they have become an integral part of people's daily lives [1][2][3]. Users can search for people who share their interests, share check-in and check-out locations, and upload photos of places they've visited. Location Based Social Networks (LBSN) encourage users to discover destinations and guide them to familiar or unfamiliar locations. An event is a scheduled occurrence that contains

spatial information. People from various countries who attend the event share their reactions to the events [4] as well as location information, resulting in location-based social media (LBSM). When compared to a social media post without geotagging, the location parameter is estimated to be twice as valuable. These reactions convey sentiment classified as positive or negative.

The objective of this work is to examine the sentiment of people who tweet from the venue. Sentiment-based marketing is possible due to this type of analysis, which extracts geolocation as well as views and interests. Marketers get access to extensive information from the posts, such as check-in statistics, activities, patterns, and transactions, all of which can be used to provide commercial recommendations. Marketers can use these tweets from events to become proactive organizers to generate or involve customers in business. Recommendations such as top-rated locations (in vicinity to event locations), activities to participate in (such as sight-seeing, dining, and shopping), and friends to collaborate can be posted to Tweeter. We use Random Forest classification to implement the Adaptive Boosting approach. Freund and Schapire presented Adaboost as one of the boosting ensemble strategies in 1996. An ensemble system, also known as a multiple classifier system, is a supervised learning model.

Data can either be balanced with an equal number of instances in both classes and imbalanced data with an unequal number of occurrences [6]. For example, fraud detection and disease detection are both examples of imbalanced data. Classification algorithms classify the majority class more accurately in comparison to minority class. In real life situations for enhancing marketing minority class prediction plays an important role. One way for reporting with an imbalance class is to 'boost' the dataset using classification methods [7][8][9]. We propose AdaBoost Ensemble Approach using Random Forest classification to address Imbalanced data.

The remainder of the paper is organized as follows: Section 2 goes over the associated work. Section 3 elaborates on the proposed Modified Boosted Random Forest (MBRF) method using an architecture model with heuristics and data sources. Section 4 depicts the studies carried out, which include i) a performance comparison of the MBRF and decision tree using performance measures also at different folds. ii) comparison of MBRF using ROC curves. Finally, in section 5 we'll make a few remarks with the conclusion.

2. Related work

LBSM has been used by researchers to analyze event usage and traveler mobility patterns. Natural disasters such as hurricanes, earthquakes, and other natural disasters are also exploited through social network systems. People actively utilize social media to share disasters by tweeting about them, sending out alerts, and marking them as safe [10][11]. We discuss how researchers have used Adaboost and Random Forest (RF) classification to classify species, traffic signs, cancer cells, crops, and many other things in the second section of related work. Random Forest is utilized in medical sector predictions, such as neuroimaging data in alzheimer's disease [12], optimal drug therapy [13], prediction of early kidney transplant [14][15], COVID-19

prediction using patient's symptoms [16]. Ensemble method-based architecture using random forest predicts employee's turn over [17], for classifying urban land cover [18], in soil data [19][20], in twitter sentiment analysis with the polarity detection task in [21][22][23].

Related work has examined imbalanced class problems and techniques which are applied is discussed in table 1.

Reference Number	Technique	Methods and Description
[24]	One-class classification techniques	using local outlier factor, one-class support vector machine and isolation forest methods. Better performances when the datasets have very high-class imbalance ratios.
[25]	SMOTE Oversampling and Random Oversampling techniques	ML algorithms such as KNN and Naïve Bayes. Increases the number of minority class data.
[26]	Im.ADABoost combining weighted -SVM	Initializes error weights and calculates confidence weights.
[27]	Tomek link under sampling algorithm.	Balancing approach - oversampling the minority class data or undersampling the majority class data.
[6]	Modified Boosted SVM (MBSVM)	distance-based weight-update rule and sign based classifier removal technique.
[28]	Wang's Boosted SVM	Cost-sensitive boosting algorithm.
[29]	Two-stage clustering-based surrogate model	Undersampling approaches.
[30]	Boosting and crossover methods real-world medical datasets	Optimizing the ratio of the two classes with different imbalance ratios.
[31]	Boosted support vector machine (B-SVM)	Classification techniques for predicting the credit ratings of banks.

Table 1. Research work on Imbalanced data and its techniques.

3. Working of MBRF

With the use of smart devices growing, social media's popularity is at an all-time high. Individuals can share their location and interests through photographs or tweets on location-based social media (LBSM). The location parameter serves higher information content when compared to social media post without geotagging. By analyzing reactions posted by attendees at global events, business and the economy can be elevated. Each reaction indicates a state of mind, emotion, or interest, along with a sentiment, all of which can be used to provide commercial recommendations. Such as travel locations, activities to participate in (such as sight-seeing, dining, and shopping), and friends to collaborate with based on similarity are some of such suggestions.

Geotagged tweets from event locations are collected and sentiment is analyzed to create a commercial recommendation system. The proposed MBRF approach has two significant changes to the MBSVM: (1) a distance-based weights updating strategy based on K Medoids and (2) RF classification based on the ensemble adaboosting approach rather than Boosted SVM.

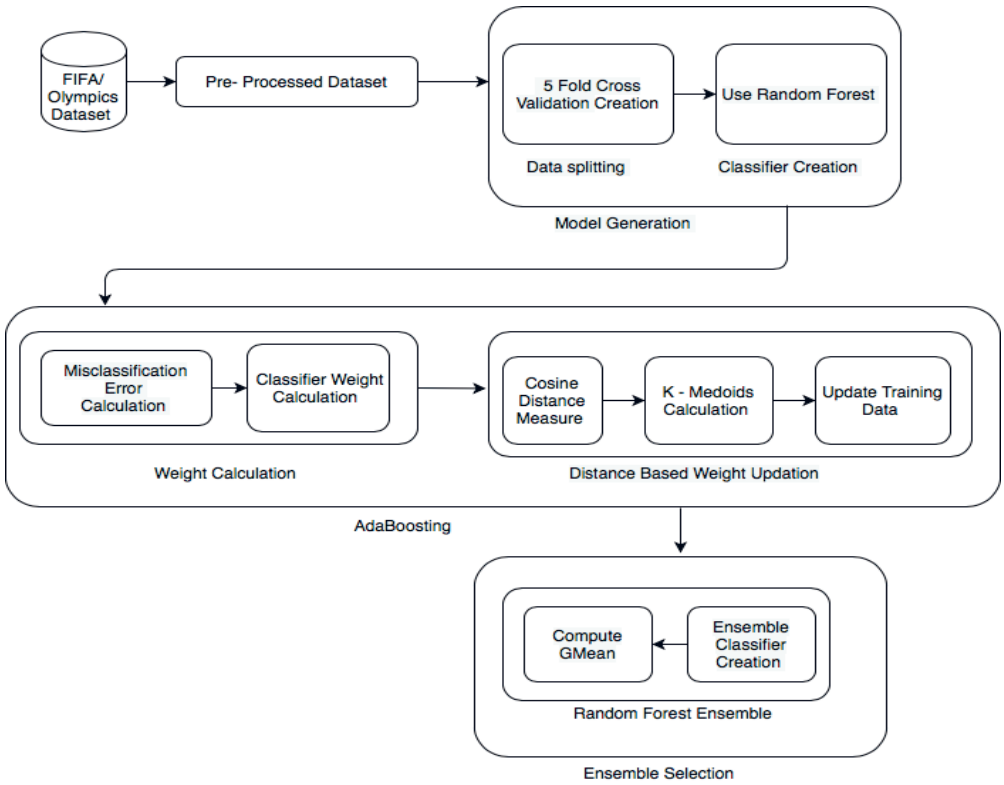

Figure 1. Data Model for MBRF

Figure 1 depicts the MBRF algorithm which has three phases: model generation, adaboosting, and ensemble selection of an MBRF data model. Adaboosting is implemented by MBRF in an iterative method in which weak learners are fitted and aggregated to the ensemble model.

3.1. Data Collection

The geotagged tweets are retrieved from Twitter using the Twitter API and annotated with hashtags listed in table 2 in English language and a sample screenshot of the tweets in figure 2.

Event Name	Hashtags
FIFA	#worldcup, #worldcupfinal, #FIFAStadium, #Russia2018,#worldcup2018, #FIFA, #WorldCupRussia
Olympics	#pyeongchang2018, #winterOlympics , #Olympics

Table 2. Hashtags used to extract tweets

Lowercase conversion, stop words, and special character removal are all performed on all retrieved tweets. The most prevalent terms in a language are referred to be stop words. Such tokens, for example, 'and', 'the', do not have a semantic meaning or sentiment attached with them. Twitter allows the use of hashtags (#) to indicate a trending topic, however special characters such as @,!,$, and so on are eliminated because they are meaningless.

Figure 2. Screenshot of the tweets after extraction process

We refer to tweets as either positive or negative observations throughout this paper. Initially, all observations are given the same weights. As a result of the learning process, the identified misclassified observations will be given higher weights [32] [33]. This step is repeated n times to create a strong classification model. A model like this can classify any new observation by predicting its classification using its weights [34]. As previously stated, MBRF employs Random Forest Classification with discrete class labels (for instance cloudy or sunny).

3.2. Location Data

Twitter API's supports tweeter to post tweet and include approximate location if they decide to enable geolocation setting with "Tweet with location". Each tweet contains the place attribute comprising of latitude and longitude or the exact location from where tweeter has posted the tweet. FIFA 2018 games were held in 11 locations and Olympics 2018 games across 10 locations as shown in table 3.

FIFA 2018	Olympics 2018
Kaliningrad	Pyeongchang-gun
Kazan	Korea
Moscow	Seoul
Nizhny Novgorod	Gangneung-si
Rostov-on-Don	Ongjin
Saint Petersburg	Wonju
Samara	Jeongseo
Saransk	Jongno
Sochi	Yangyeong
Volgograd	Jung-gu
Yekaterinburg	-

Table 3. Locations of FIFA and Olympics Events

3.3. Description of datasets

The data set is separated into two sets named as training sets, which is used to form the learned hypothesis, and a separate validation set, which is used to evaluate the accuracy. The description of both datasets is given in Table 4.

Characteristics	FIFA	Olympics
Number of Observations	25,000	28,000
Number of variables	5	5
Positive Class	19,000	23,000
Negative Class	3000	5000
Imbalanced Ratio	3.1666	4.6

Table 4. Imbalanced dataset and their characteristics

3.4. Phases of MBRF

MBRF algorithm has four phases:
Phase 1: Model Generation

(i) Data splitting: The algorithm starts by splitting the data using 5-fold cross validation f = 1,2…5. For each fold we have one part serving as validation data ($Xval_f$) and the rest as training data ($Xtrain_f$). The validation and training parts are switched until all five components have been validated at least once.

(ii) Classifier Creation:
Initial weights $W_i = \frac{1}{N}$ where i = 1, 2,….., N ….. Equation (1)
N is the number of tweets or also called as observations.

We have classifier RF_{mf} created for all the folds i.e number of classifiers created is RF_{m1} ,RF_{m2} ,….,RF_{m5}.Random Forest (RF) is modelled by setting the hyperparameters.

RF_{mf} = RFTrain($Xtrain_f$, 'method')
$RF_{mf} = [w_1, w_2, \ldots w_n]$

Where m is the current iteration, m =1,2,….M, M is the Maximum number of iterations.

Phase 2: Adaboosting

(i) Misclassification Error (e_{mf}): Misclassification is computed by classifying $Xtrain_f$ using RF_{mf}

$$e_{mf} = \frac{\Sigma W_{if} * Prediction\_error\_rate_{if}}{\Sigma w_{if}} \ldots. \text{ Equation (2)}$$

The $Prediction\_error\_rate_{if}$ is calculated for each training observation 'i' across all folds (f). If an observation is misclassified 'i' =0, otherwise it is assigned a value of 1. Hence, we have misclassification error for each fold $e_{m1}, e_{m2} \ldots e_{m5}$.

(ii) The component classifier (α_{mf}) the classifier weight is derived from the misclassification error as shown in Equations 3. In this case,

$$\alpha_{mf} = log \frac{1-e_{mf}}{e_{mf}} \quad \text{if} \quad emf < 0.5 \ldots. \quad \text{Equation (3a)}$$

$$\alpha_{mf} = 0 \quad \text{if} \quad emf > 0.5 \ldots. \quad \text{Equation (3b)}$$

$$\alpha_{mf} = \alpha_{mf} + \varepsilon \quad \text{if} \quad em = 0 \ldots. \quad \text{Equation (3c)}$$

When the error e_{mf} becomes 0 we assume to add a small value ε, a constant value used to adjust the magnitude of the component classifier α_{mf}(where ε = 0.001)

Observation weight update (w_{if}): we change the weights of all misclassified observations identified in the phase 2 and transform the data to classify them correctly.

Phase 3: Ensemble Selection

In this phase, for every iteration 'm', we create component classifier (G_{mf})

$$G_{mf} = \alpha_{mf} * S_{mf} \quad \text{…. Equation (4)}$$

Where S_{mf} is the base classifier.

Next, we compute the various evaluation metrics for each fold's ensemble classifier using the respective fold's Xval,f.

Elaborating MBRF algorithm in different phases:

Phase 1

We create a data Matrix X with N observations which includes $X_{train,f}$ and $X_{val,f}$. to generate Y a class label of data Matrix X. Tweet, Location, Sentiment, Number of Followers, and Hashtags are used as attributes of the X. Breiman invented the Random Forest (RF) learning method, which is an ensemble of Classification and Regression Trees (CART) procedures used to build a forest.

The RF uses random variation to build a forest, many individual decision trees are generated using the $X_{train,f}$. This tree construction considers all the attributes the data Matrix X, tree is grown and aggregated as base learners for prediction. In general, higher number of trees implies more resilient forest. The prediction of the model emphasizes on aggregating the results generated from different decision trees. RF can be trained on a dataset that consists of a collection of tree predictors, each of which is based on a randomly sampled vector.

The best prediction of accuracy results of the MBRF model can derived by the tuning process of hyperparameters. MBRF hyperparameters are set using Random Forest classifier built with Scikit-learn (Sklearn) a useful and robust library for machine learning in Python. We elaborate the Random Forest hyperparameters settings in table 5 set prior to the learning process.

Hyperparameter	Description
Estimators	The number of trees in the MBRF model. We've set n estimators to 120 and 90 estimators for the FIFA and Olympics events, respectively.
Criteria	splitter object whose function is to discover the best internal node split. Assesses the goodness of a split and to rank the importance of the qualities.
Maximum Features	Influences tree diversity. *max_features* = 1 creates forests with diverse and complicated trees; *max_features is* close to the number of features, it creates forests with simpler trees.
Maximum Depth	*max_depth* = None by default, which means that each tree will extend until every leaf is pure. Otherwise, splitting occurs until all the leaves are pure, indicating that all the leaves belong to the same.

Table 5. Hyperparameters settings in MBRF Model

The number of trees to construct the Random Forest was assessed using the out-of-bag (OOB) error rate. OOB samples are defined as a set of bootstraps (produced by random resampling from the training dataset) that are not part of the dataset but are used as a test dataset. OOB error also known as out-of-bag estimate, is used to estimate the random forest prediction error. The critical part in the feature selection process is selection of the optimal number of features determined based on the minimum OOB error rate. OOB prediction computes classification error for each 'n' observations by taking the majority vote for the i^{th} observation. For the classification, max features set to 'sqrt' to examine the OOB error rate versus n estimators. Figure 3 depicts the out-of-bag error rate with n estimators for event datasets after the MBRF is applied. The results clearly show that FIFA and the Olympics have the lowest OOB error rates at n estimators, at 120 and 90, respectively.

Once the model is fitted computation of error (e_{mf}) for misclassified data is estimated by weighted average method. Here misclassified observations are weighed focusing on weak learners of $X_{train,f}$. The weighted average method detects and boosts the misclassified observations.

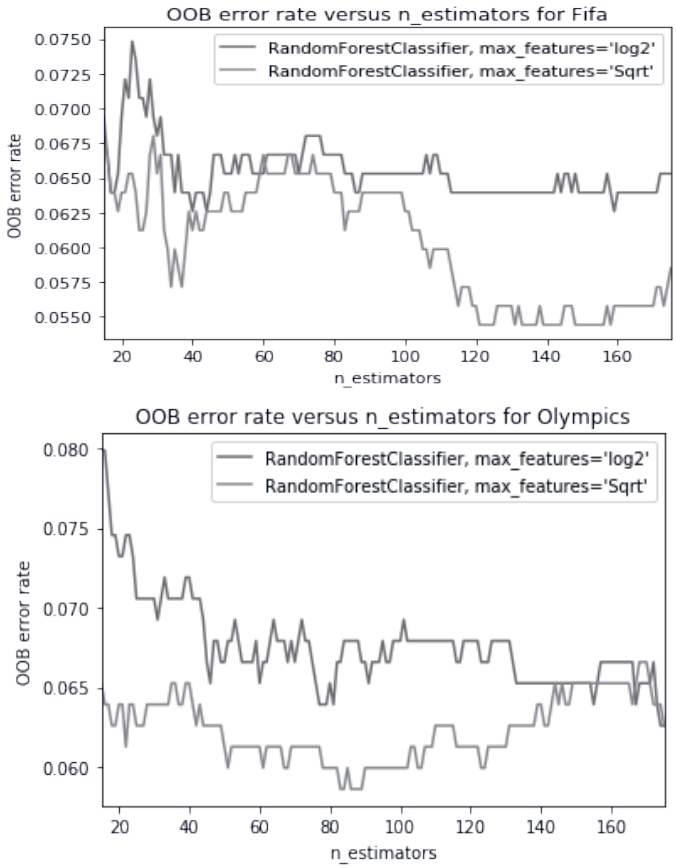

Figure 3. Demonstration of Out of the bag error rate with n_estimators for events datasets

Phase 2

The weights of misclassified observations are updated using the distance ratio and classifier weight (mf). Weight update procedure prioritizes the distance of each misclassified observation from its own group's centroid and handles them separately. The weight of these observations is updated at each iteration through clustering to correctly classify them using distance measurement.

The weight update procedure use K medoids to select one of the points in the cluster as centroid [35][36] to similar observations are grouped together in a cluster, while dissimilar observations are grouped together in other clusters [37][38]. Accordingly, we have calculated the centroid of both groups namely group1 and group2 (since k =2). Each misclassified observation in group1 is compared with distance from centroid 1 and 2. If the data instance's distance is far away from centroid 1.

The below expression is used for weight update using the below equation 5.

$$w_{if} = w_{if} * \exp\left(\alpha_{mf} * \frac{d_{i2f}}{d_{i1f}} * \ prediction\_error\_rate_i \right) \ if\ d_{i1f} > d_{i2f}\ for\ group\ 1 \quad \text{…. Equation (5a)}$$

$$w_{if} = w_{if} * \exp\left(-\alpha_{mf} * \frac{d_{i1f}}{d_{i2f}} * \ prediction\_error\_rate_i \right) \ if\ d_{i2f} > d_{i1f}\ for\ group\ II \quad \text{…. Equation (5b)}$$

We use equation 5a to make the data instance closer to group1 centroid such that observation shall be classified correctly. The same implies to each misclassified observation in group 2, the distance from group 2 centroid is compared. If the data instance's distance is far away from centroid 2. Equation 5b is applied to make data instances closer to group 2 centroid such that observation shall be classified correctly.

As previously stated, our research focuses on the events dataset, which is primarily a social interaction of user tweets in the form of text [39][40]. The cosine of the angle between observations is measured by cosine similarity, which is a measure of similarity between two non-zero vectors of an inner product space [41]. As shown in equation 6, cosine similarity is used to calculate the distance between the misclassified observations A and B.

$$A.\ B = ||A||\ ||B||\ \text{Cos}\ \theta \quad \text{…. Equation (6)}$$

Based on the clustering we have shown the locations plotted as positive and negative in figure 4. These results are identified with K medoids process with group1 and group2 to select one of the points in the cluster as centroid.

Phase 3: Ensemble Selection

The Random Forest Ensemble stage of MBRF is where the Ensemble selection processing takes place. Ensemble pruning, also known as post-selection, is used in this process. During this process, a subset of the original ensemble's base

classifiers/learners is removed with no loss of predictive performance. This pruning allows an ensemble to grow freely before pruning it, resulting in an effective ensemble [42][43][44]. Finally, we compute the ensemble performance of each fold using various evaluation metrics such as precision, recall, accuracy widely used for recommendation algorithms in equations 7-10 [45][46].

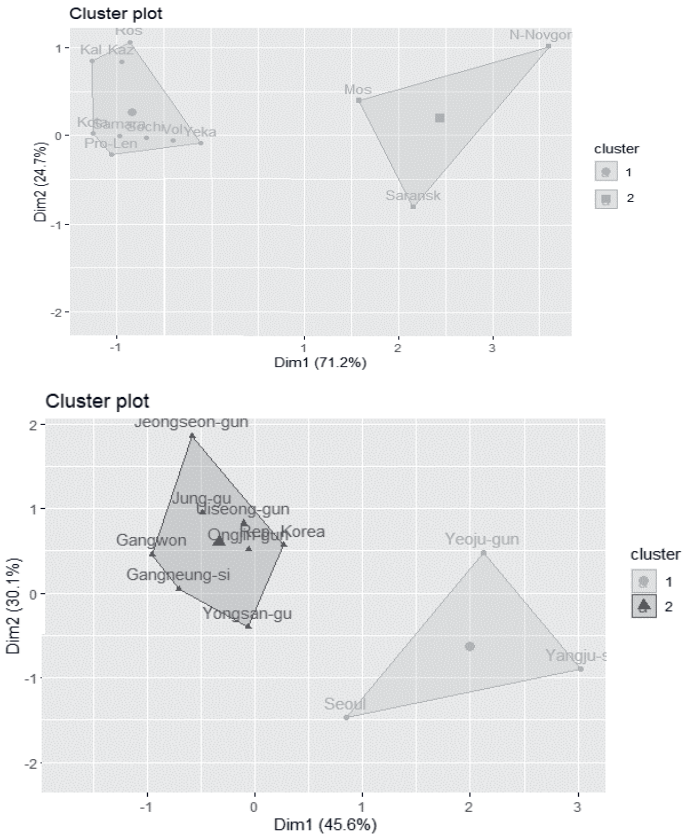

Figure 4. Location Based Clustering with K=2 at FIFA and Olympics events

We define the recommendation metrics which were applied to evaluate the MBRF model and DTC.

a. Precision (P) the fraction of recommended items that is relevant to the user.

$$P = \frac{N_{rs}}{N_s} \quad \text{.... Equation (7)}$$

b. Recall (R) the fraction of relevant items that are also part of the set of recommended items.

$$R = \frac{N_{rs}}{N_r} \quad \text{…Equation (8)}$$

Where

N_{rs} – Correctly recommended locations

N_s – Total recommended locations
N_r – Total relevant locations

c. Accuracy (A) is the fraction of correct recommendations out of total possible recommendations.

$$A = \frac{N_{rc}}{N_{rp}} \quad \text{.... Equation (9)}$$

N_{rc} -Number of correct recommendations
N_{rp} – Total possible recommendations

d. F1 – Score the harmonic mean of precision and recall is considered as in equation 10.

$$F1 - Score = 2\ \frac{P.R}{P+R} \quad \text{.... Equation (10)}$$

4. Results

4.1. Performance Comparison using performance metrics.

Various metrics – precision, recall, F1- score, accuracy and Support is measured using Random Forest Classifier by setting the splitter criterion as 'Gini' and 'Entropy'. Entropy is a measurement of impurity or randomness in observations. The degree of impurity is 0 -1 order. Gini Index calculates the amount of probability of a specific feature that is classified incorrectly when selected randomly. We compare the performance of MBRF with DTC using splitter criterion set as 'Gini' and 'Entropy'. Table 6 has the comparison results tabulated and it is evident that MBRF performs better than DTC. These findings also show that MBRF outperforms DTC using the 'Gini' and 'Entropy' criteria. In table 6 we attempt to bring out the accuracy of each fold using MBRF and DTC to prove MBRF to be better than DTC across all the folds.

Classifier Name	Event	Precision	Recall	F1-score	Support	Accuracy
MBRF (criterion= 'Gini')	FIFA	0.92	0.93	0.93	303	93.36
	Olympics	0.90	0.92	0.91	206	91.291
MBRF (criterion= 'Entropy')	FIFA	0.92	0.94	0.93	303	93.99
	Olympics	0.91	0.90	0.90	206	94.747
DTC (criterion= 'Gini')	FIFA	0.91	0.91	0.91	303	91.419
	Olympics	0.90	0.90	0.90	206	90.29
DTC (criterion = 'Entropy')	FIFA	0.90	0.92	0.92	303	91.749
	Olympics	0.90	0.90	0.90	206	90.56

Table 6. Performance Metrics for the Events

Classifier Name	Accuracy	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
MBRF(criterion= 'Entropy')	FIFA	93.01	93.44	92.76	93.79	93.24
	Olympics	94.46	94	94.12	94.55	94
DTC (criterion ='Entropy')	FIFA	91.58	91.16	91.49	91	91.65
	Olympics	91.23	90.11	90.4	90.56	91.36

Table 7. Performance Metrics for the Events across the folds

4.2. Comparison of classification performance through ROC curve

One of the most common methods for comparing binary classification models is Receiver Operating Characteristic curve (ROC). The ROC analysis is demonstrated using the ROC curve, which is a tool for evaluating, comparing, and selecting the best classifier based on classification performance. The False Positive Rate (FPR) and True Positive Rate (TPR) for both events are plotted using a ROC curve. The terms FPR and TPR are derived from the confusion matrix, as shown in equations 11 and 12, respectively. Where TP rate refers to the percentage of correctly identified positives. The TPR rate is also known as the sensitivity or recall rate.

$$TPR = \frac{TP}{TP+FN} \qquad \text{…. Equation (11)}$$

The proportion of correctly identified negatives is measured by the TN rate. TNR rate is also known as specificity number of true positives - TP, number of true negatives - TN, number of false positives - FP, and number of false negatives – FN, where the total number of observations is equal to the sum of TP + TN + FP + FN = n.

$$TNR = \frac{TN}{TN+FP} \qquad \text{…. Equation (12)}$$

The geometric mean (Gmean) metric, formulated in equation 13, is an ensemble classification for measuring sensitivity and specificity.

$$Gmean = \sqrt{TPR * TNR} \qquad \text{… Equation (13)}$$

We created the ROC curves for both datasets using the framework described above. Figure 5 shows the ROC curve for FIFA and Olympics events to demonstrate MBRF. On both datasets, we calculated AUC using MBRF values. On the FIFA dataset, the average AUC value is 0.96. On both datasets, we calculated AUC using MBRF values. The average AUC value on the FIFA dataset is 0.96, and the Olympics is 0.97, indicating that our proposal will be performed well.

5. Conclusion

Ensemble methods are highly effective at boosting performance of simple learners. They have earlier been used in many applications; in this work we apply on analysis

of location based social networks for events. Random Forest (RF) is a stable ensemble of trees with a self -testing feature that is robust to overfitting. From the results, MBRF proves to be a good classifier to predict reactions of people's events and recommend them suitably based on current need. We have provided empirical evidence with classical measurement metrics yielding good results (accuracy, precision and recall) for classification results. Limitation/future enhancement - We have focused only on two global events we can extract tweets of other events such as awards and recognition, musical concerts. Further we can consider other social media such as Instagram and Facebook rather than using only tweets.

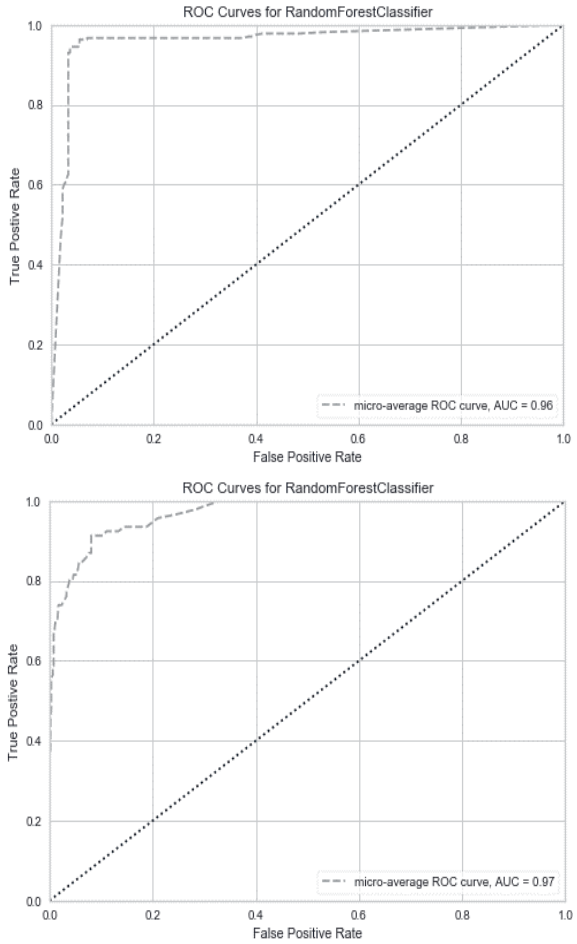

Figure 5. Demonstration of ROC curves for FIFA and Olympics Events

References

[1] Cao, Xiaodong et al. "Using Twitter to Better Understand the Spatiotemporal Patterns of Public Sentiment: A Case Study in

Massachusetts, USA.” International Journal of Environmental Research and Public Health 15 (2018)

[2] Molitor, Dominik, et al. "Effectiveness of location-based advertising and the impact of interface design." Journal of Management Information Systems 37.2 (2020): 431-456.

[3] Le, Xuan Cu, and Tran Hung Nguyen. "A framework of location-based advertising effectiveness: perspectives of perceived value and satisfaction." Asian Journal of Business Research Volume 11.3 (2021).

[4] Alexander Dunkel, Gennady Andrienko, Natalia Andrienko, Dirk Burghardt, Eva Hauthal & Ross Purves. “A conceptual framework for studying collective reactions to events in location-based social media”.– International Journal of Geographical Information Science Vol. 33, (2019), Issue 4

[5] Wang, Changbai, Shuzhan Xu, and Junxin Yang. "Adaboost Algorithm in Artificial Intelligence for Optimizing the IRI Prediction Accuracy of Asphalt Concrete Pavement." Sensors 21.17 (2021): 5682

[6] Sundar R., Punniyamoorthy M , “Performance enhanced Boosted SVM for Imbalanced datasets”, Applied Soft Computing , Elsevier, Volume 83, October 2019, 105601

[7] Wang, Wenyang, and Dongchu Sun. "The improved AdaBoost algorithms for imbalanced data classification." Information Sciences 563 (2021): 358-374.

[8] Jiang, Xue, et al. "An imbalanced multifault diagnosis method based on bias weights AdaBoost." IEEE Transactions on Instrumentation and Measurement 71 (2022): 1-8.

[9] Li, Xiao, and Kewen Li. "Imbalanced data classification based on improved EIWAPSO-AdaBoost-C ensemble algorithm." Applied Intelligence 52.6 (2022): 6477-6502.

[10] Ullah, Irfan, et al. "RweetMiner: Automatic identification and categorization of help requests on twitter during disasters." Expert Systems with Applications 176 (2021): 114787.

[11] Jain, Ankit Kumar, Somya Ranjan Sahoo, and Jyoti Kaubiyal. "Online social networks security and privacy: comprehensive review and analysis." Complex & Intelligent Systems 7.5 (2021): 2157-2177.

[12] Tripathi, Satvik, Alisha Isabelle Augustin, and Edward Kim. "Longitudinal Neuroimaging Data Classification for Early Detection of Alzheimer’s Disease using Ensemble Learning Models." (2022).

[13] Guo, Wei, et al. "A machine learning model to predict risperidone active moiety concentration based on initial therapeutic drug monitoring." Frontiers in psychiatry 12 (2021).

[14] Seyahi, Nurhan, and Seyda Gul Ozcan. "Artificial intelligence and kidney transplantation." World Journal of Transplantation 11.7 (2021): 277.

[15] Thongprayoon, Charat, et al. "Feature Importance of Acute Rejection among Black Kidney Transplant Recipients by Utilizing Random Forest Analysis: An Analysis of the UNOS Database." Medicines 8.11 (2021): 66.

[16] Soui, Makram, et al. "NSGA-II as feature selection technique and AdaBoost classifier for COVID-19 prediction using patient's symptoms." Nonlinear dynamics 106.2 (2021): 1453-1475.

[17] Gao, Xiang, Junhao Wen, and Cheng Zhang. "An improved random forest algorithm for predicting employee turnover." Mathematical Problems in Engineering 2019 (2019).

[18] Zhang, Tianxiang, et al. "Sentinel-2 satellite imagery for urban land cover classification by optimized random forest classifier." Applied Sciences 11.2 (2021): 543.

[19] Rani, Alka, et al. "Identification of salt-affected soils using remote sensing data through random forest technique: a case study from India." Arabian Journal of Geosciences 15.5 (2022): 1-16.

[20] Lkeagwuani, Chijioke Christopher. "Estimation of modified expansive soil CBR with multivariate adaptive regression splines, random forest and gradient boosting machine." Innovative Infrastructure Solutions 6.4 (2021): 1-16.

[21] Bibi, Maryum, et al. "A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis." Pattern Recognition Letters 158 (2022): 80-86.

[22] Onan, Aytug. "Ensemble of classifiers and term weighting schemes for sentiment analysis in Turkish." Scientific Research Communications 1.1 (2021).

[23] Carvalho, Jonnathan, and Alexandre Plastino. "On the evaluation and combination of state-of-the-art features in Twitter sentiment analysis." Artificial Intelligence Review 54.3 (2021): 1887-1936.

[24] Tsai, Chih-Fong, and Wei-Chao Lin. "Feature selection and ensemble learning techniques in one-class classifiers: an empirical study of two-class imbalanced datasets." IEEE Access 9 (2021): 13717-13726.

[25] Hayaty, Mardhiya, Siti Muthmainah, and Syed Muhammad Ghufran. "Random and synthetic over-sampling approach to resolve data imbalance

in classification." International Journal of Artificial Intelligence Research 4.2 (2020): 86-94

[26] Quang, Vo Duc, Tran Dinh Khang, and Nguyen Minh Huy. "Improving ADABoost Algorithm with Weighted SVM for Imbalanced Data Classification." Future Data and Security Engineering: 8th International Conference, FDSE 2021, Virtual Event, November 24–26, 2021, Proceedings 8. Springer International Publishing, 2021.

[27] Kamaladevi, M., V. Venkataraman, and K. R. Sekar. "Tomek link undersampling with stacked ensemble classifier for imbalanced data classification." Annals of the Romanian Society for Cell Biology (2021): 2182-2190.

[28] Zhang, Wei, et al. "Cost-sensitive multiple-instance learning method with dynamic transactional data for personal credit scoring." Expert Systems with Applications 157 (2020): 113489.

[29] Le, Hoang Lam, et al. "EUSC: A clustering-based surrogate model to accelerate evolutionary undersampling in imbalanced classification." Applied Soft Computing 101 (2021): 107033

[30] Desuky, Abeer S., Asmaa Hekal Omar, and Naglaa M. Mostafa. "Boosting with crossover for improving imbalanced medical datasets classification." Bulletin of Electrical Engineering and Informatics 10.5 (2021): 2733-2741.

[31] Viswanathan, Hari Hara Krishna Kumar, et al. "A modified boosted support vector machine to rate banks." Benchmarking: An International Journal 28.1 (2021): 1-27.

[32] Sevinç, Ender. "An empowered AdaBoost algorithm implementation: A COVID-19 dataset study." Computers & Industrial Engineering 165 (2022): 107912.

[33] Sandica, Ana-Maria, and Alexandra Fratila. "Implications of macroeconomic conditions on Romanian portfolio credit risk. A cost-sensitive ensemble learning methods comparison." Economic Research-Ekonomska Istraživanja (2021): 1-20.

[34] Hornung, Roman. "Diversity forests: Using split sampling to enable innovative complex split procedures in random forests." SN computer science 3.1 (2022): 1-16.

[35] Guan, Renchu, et al. "Deep feature-based text clustering and its explanation." IEEE Transactions on Knowledge and Data Engineering (2020).

[36] Kumar, Jay, et al. "An online semantic-enhanced Dirichlet model for short text stream clustering." Proceedings of the 58th annual meeting of the association for computational linguistics. 2020.

[37] Budiaji, Weksi, and Friedrich Leisch. "Simple K-medoids partitioning algorithm for mixed variable data." Algorithms 12.9 (2019): 177.

[38] Oktarina, Cahyani, Khairil Anwar Notodiputro, and Indahwati Indahwati. "Comparison of k-means clustering method and k-medoids on twitter data." Indonesian Journal of Statistics and Its Applications 4.1 (2020): 189-202.

[39] Almaslukh, Bandar. "Forensic analysis using text clustering in the age of large volume data: A review." International Journal of Advanced Computer Science and Applications 10.6 (2019).

[40] Rodriguez, Mayra Z., et al. "Clustering algorithms: A comparative approach." PloS one 14.1 (2019): e0210236.

[41] Alqaisi, Rana, Wasel Ghanem, and Aziz Qaroush. "Extractive multi-document Arabic text summarization using evolutionary multi-objective optimization with K-medoid clustering." IEEE Access 8 (2020): 228206-228224

[42] Zhang, Hua, et al. "Two-stage bagging pruning for reducing the ensemble size and improving the classification performance." Mathematical Problems in Engineering 2019 (2019).

[43] Ahmed, Ahmed Mohamed, Ahmet Rizaner, and Ali Hakan Ulusoy. "A novel decision tree classification based on post-pruning with Bayes minimum risk." PLoS One 13.4 (2018): e0194168.

[44] Iorio, Carmela, et al. "Informative trees by visual pruning." Expert Systems with Applications 127 (2019): 228-240.

[45] F.O. Isinkaye, Y.O. Folajimi, B.A. Ojokoh, "Recommendation systems: Principles, methods and evaluation , Egyptian Informatics Journal", Volume 16, Issue 3, November 2015, Pages 261-273

[46] Jie Bao, Yu Zheng, David Wilkie, Mohamed F Mokbel, "Recommendations in location-based social networks: a survey", Geo Informatica, July 1 2015