MiNB: Minority Sensitive Naïve Bayesian Algorithm for Multi-Class Classification of Unbalanced Data

Pratikkumar Barot
Computer Engineering Department, Gujarat Technological University, India
pratikabarot@gmail.com

Harikrishna Jethva
Computer Engineering Department, Gujarat Technological University, India
hbjethva@gmail.com

Abstract: *The unbalanced nature of data makes it tough to achieve the desire performance goal for classification algorithms. The sub-optimal prediction system isn't a viable solution due to the high misclassification cost of minority events. Thus accurate imbalanced data classification could be a path changer for prediction in domains like medical diagnosis, judiciary, and disaster management systems. To date, most of the existing studies of imbalanced data are for the binary class dataset and supported by data sampling techniques that suffer from loss of information and over-fitting. In this paper, we present the modified naïve Bayesian algorithm for unbalanced data classification that eliminates the requirement of data level sampling. We compared our proposed model with the data sampling technique and cost-sensitive techniques. We use minority sensitive TP Rate, class-specific misclassification rate, and overall performance parameters such as accuracy, f-measure and G-mean. The result shows that our proposed algorithm shows a more optimal result for unbalanced data classification. Results shows reduction in misclassification rate and improve predictive performance for the minority class.*

Keywords: *Imbalanced data learning, weighted naïve bayesian, cost-sensitive learning, multi-class unbalanced data.*

Received October 13, 2020; accepted December 13, 2021
https://doi.org/10.34028/iajit/19/4/5

1. Introduction

An unbalanced data learning is a classification approach that performs accurate classification of unbalanced data [7]. Unbalanced data has an uneven distribution of the data [4]. If the unbalanced dataset has only two classes then it is called a binary-class unbalanced dataset and if it has more than two classes then it is called a multi-class unbalanced dataset. An unbalanced dataset may have either a between-class imbalance or within-class imbalance or both [21]. Between-class unbalanced datasets have an uneven distribution of class labels. It has a positive class (minority) and negative class (majority) instances. Majority class instances dominate the minority class instances during the learning process [17, 18].

The Imbalance Ratio (IR) indicates the level of imbalance between the classes [21]. For the binary-class dataset, the imbalanced ratio is the ratio of majority class instances and minority class instances. As per Mujalli *et al*. [17], if minority instances are lesser than 35% of total instances then it is considered as an imbalance dataset. The within-class unbalanced dataset shows the uneven distribution of sub-concepts within the class.

Unbalanced data make traditional classification algorithm less optimal. Classification of the multi-class unbalanced dataset is a more challenging task to perform [10]. The techniques proposed so far for the unbalanced data classification are of mainly two types: data level techniques [8, 9] and algorithm level techniques [2, 15].

Data level technique involves the sampling of the data that either replicate the minority class instances or remove the majority class instances to balance the dataset [18]. Oversampling and under-sampling are two widely used data level techniques to handle the unbalanced nature of the data [1]. In the oversampling, new minority instances are created either by replicating existing minority samples or by creating new synthetic minority samples. In the under-sampling, the majority class instances are removed from the dataset [25].

Algorithm level techniques require the development of a new algorithmic approach or modification of the existing algorithm to handle unbalanced data. One class learning, cost-sensitive learning, and ensemble-based method are some of the techniques used for algorithmic approach [10, 21]. The algorithmic approach is more accurate as compared to the logistic regression models [16] and the sampling techniques [13].

Traditional classification algorithms with accuracy above 90% are not the optimal approach for the unbalanced data. Because they favor the majority class and thus incorrectly classify the minority class instances [1, 11, 21]. Most of the existing methods of unbalanced classification are designed for binary class problems [4, 22, 26], and they use under-sampling for majority class and/or oversampling for the minority class. The sampling techniques have some drawbacks like over-fitting, increased computational time, and loss of information [9, 13, 29]. Classification of the absolute minority class is a more challenging task because of the little representation of those instances within the dataset [7, 21].

As per Stefanowski [21], for the relatively unbalanced dataset, it is possible to collect more minority examples and increase the size of the dataset by maintaining the imbalance ratio. In this case, the absolute cardinality of minority class is no longer rare and it is easier to classify. But absolute minority class instances are very rare and the sampling method may create a large dataset to make them noticeable. This restricts the use of the sampling method for the absolute minority class instances.

Cost-sensitive learning has good performance as compared to the sampling methods [7, 29]. Effective cost-sensitive learning for imbalanced data classification is in demand and recently many authors proposed this as a future work [24, 31]. Cost-sensitive learning aims to reduce either the test cost or the misclassification cost or both. The misclassification cost is a cost incurred due to the classification error [28]. In the domains like medical diagnosis and the judicial system, there is a huge misclassification cost is associated with the misclassification of the minority class.

In this paper, we proposed cost-sensitive learning for a multi-class unbalanced dataset using the modified Naïve Bayesian algorithm called Minority sensitive Naïve Bayesian (MiNB). The MiNB uses a causal relationship based feature weights. It improves the result of the unbalanced data classification and thus reduces the misclassification cost.

The remainder of this paper is organized as follows, section 2 gives information of datasets used. Section 3 discusses work related to our study. Section 4 explains our proposed MiNB model. Section 5 gives details of evaluation parameters, and section 6 discusses method implementation and comparison of results.

2. Datasets

We proposed the MiNB for a multi-class unbalanced dataset. We select the multi-class unbalanced datasets for our experiment. Table 1 shows the dataset information. We have used six multi-class datasets available in the Knowledge Extraction based on Evolutionary Learning (KEEL) repository.

Table 1. Description of the multi-class datasets.

Name	#Attributes	#Class	#Instance
Hayes-Roth	4	3	132
New-Thyroid	5	3	215
Ecoli	7	8	336
Pageblocks	10	5	548
Yeast	8	10	1484
Wine Quality	12	7	6497

3. Related Work

Patel and Thakur [18], proposed an optimally weighted fuzzy based nearest neighbor strategy for the unbalanced data. They optimized the K-Nearest Neighbor (KNN) algorithm by selecting the different values of K for different classes. They found that their proposed model works well as compared to traditional fuzzy based technique.

The cost-sensitive unbalanced data learning problem resides among the top ten challenging problems [12, 18]. Recently many studies have been performed in the unbalanced data classification using the naïve Bayesian classifier [1, 17, 22, 31]. The Support Vector Machine (SVM) and KNN experiences a sudden drop in their performance if the imbalance ratio and overlapping increases [21]. The conventional decision tree-based algorithms have natural biasing towards the majority class because of the use of the information gain and the Gini index [19].

Vluymans *et al*. [27] Stated that the main drawback of the SVM is high computational cost. The C4.5-decision tree algorithm is somewhat greedy and uses an entropy-based top-down divide and conquer method and because of this, it is not capable of correct classification of minority instances [19]. The noticeable benefit of the Bayesian classifiers over the SVM and KNN based classifiers is that the former implicitly produces interpretable confidence values in the form of class membership probability estimates for each classification it makes [27, 30].

Many studies have been performed to improve the performance of the naïve Bayesian algorithm for balanced data using feature weight [19, 23, 28, 30]. However, these methods are not optimal for an unbalanced dataset. Ratnanmahatana and Gunopulos [19], uses C4.5 to select attributes for the naïve Bayesian. They used only those attributes which appear in the C4.5 decision tree. This method is not suitable for the minority class classification as it concentrates only on the majority class and ignores the minority class. Bashir *et al*. [3], proposed feature selection based on maximum likelihood logistic regression for imbalanced data of software detection.

Chomboon *et al*. [7], concluded that feature selection with oversampling is the best method for unbalanced data classification. Kong *et al*. [12] propose the improved version of a selective Bayesian classifier called Test Cost-Sensitive Naïve Bayesian (TCSNB). In TCSNB, authors select the attributes which improve the accuracy and then they remove the attributes which have high test cost. Just removal of attributes which have high test cost ultimately result in information loss and lead to the poor predictive accuracy of the minority class.

Mujalli *et al*. [17], proposed a Bayes classifier with sampling methods for unbalanced accident datasets. They predict the severity of an accident as a slight injury or major injury from input circumstances at the time of the accident. The major injury is in minority.

Zhang *et al*. [32] proposed sampling-based ensemble techniques for unbalanced data classification. They use under-sampling, re-sampling with random feature selection. The random selection of feature space does not work for real-world problems. In the same year, Triguero *et al*. [24] propose a Random Oversampling and Evolutionary Feature Weighting for Random Forest (ROSEFW-RF) algorithm for extremely unbalanced big data bioinformatics problems. The ROSEFW-RF uses oversampling as a pre-processing step which increases the size of the dataset and learning time. In their study, Triguero *et al*. [24] realize the requirement for the evolutionary feature selection approach for the unbalanced data.

Braytee *et al*. [4] proposed a cost-sensitive learning strategy for feature extraction. They proposed Cost-Sensitive Principal Component Analysis (CSPCA) and Cost-Sensitive non-Negative Matrix Factorization (CSNMF) methods for handling feature extraction from unbalanced data. The authors mention a multi-label classification problem for future work.

The Synthetic Minority Oversampling Technique (SMOTE) [6] is one of the trendsetter sampling technique for the unbalanced data. Saez *et al*. [20], proposed SMOTE-Iterative Partitioning Filter (SMOTE-IPF) that is based on the iterative-partitioning filter for the noise reduction from the unbalanced dataset. However, as per Jiang *et al*. [11], SMOTE uses the same sampling rate for all the minority class instances. So they propose a Genetic Algorithm based SMOTE (GASMOTE) which uses different sampling rate for different instances.

Herna *et al*. [10] proposed the SMOTE and Cluster Under-sampling Technique (SCUT) sampling technique. SCUT is a sampling method that combines oversampling and under-sampling. It uses SMOTE for oversampling and the Expectation-Maximization (EM) clustering method for under-sampling. The EM clustering ensures that no sub-concept representation is lost during the under-sampling.

Trisanto *et al*. [25], proposed an under-sampling and feature reduction based approach for the unbalanced data of credit card fraud detection. The authors proposed feature reduction using correlation coefficient and principal component analysis. Authors also accept that two-stage feature reduction doesn't perform well for the unbalanced credit card fraud dataset. Cost-sensitive unbalanced learning without data level sampling or feature reduction can avoid information loss and degradation of the performance.

4. MiNB

The MiNB is based on the philosophy that most of the outcomes have some exclusively responsible patterns. This theory is also statistically proved by Barot *et al*. [2] in a statistical study. They performed a detailed study and proved that some exclusive causes are strongly related to the target label. They identified class-specific most relevant features-value pairs as responsible patterns. The MiNB uses the derived pattern-base to weight the class prediction probability.

4.1. MiNB Model

Figure 1 shows the flowchart of the MiNB algorithm. The MiNB has four phases:

1. Data pre-processing and discretization.
2. Extraction of class-specific exclusive patterns.
3. Generation of pattern-base.
4. Classification and performance evaluation.

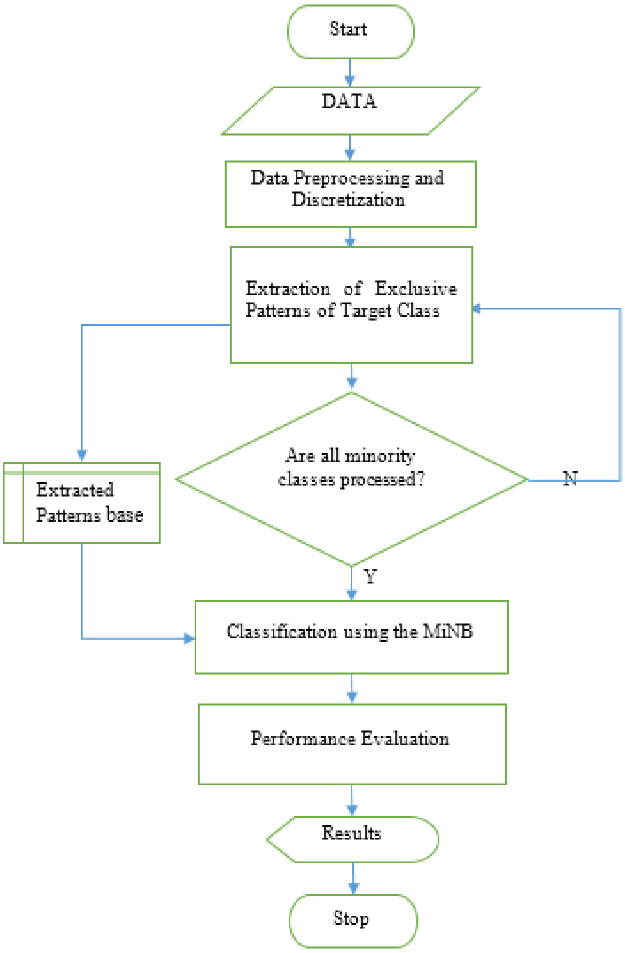
Start
DATA
Data Preprocessing and Discretization
Extraction of Exclusive Patterns of Target Class
Extracted Patterns base
Are all minority classes processed?
N
Y
Classification using the MiNB
Performance Evaluation
Results
Stop

Figure 1. Flow of MiNB algorithm.

4.2. Detail Steps of MiNB

Data pre-processing is used to handle the missing values and noise. All datasets are noise-free while missing values are filled with the attribute mean value. After handling the data level difficulties we convert continuous values into discrete values during the discretization phase.

In the next step, exclusively responsible patterns are systematically discovered and extracted for each target class. We used a modified Apriori algorithm proposed by Barot *et al*. [2] for this purpose. Feature(s) which are tightly associated with the target class are identified with their level of bonding with the target class. The causal relationship is represented as shown in Equation (1).

$$\{Ai, ..Ak, ..Ap\} => C_i \text{ [confidence=\%c]} \qquad (1)$$

Here, $\{Ai, ..Ak, ..Ap\}$ is the set of attribute-value pairs and Ci is the ith class. The confidence value indicates the strength of bonding between features and target class. If confidence value is equal to one then it means patterns are exclusive to the target class. Table 2 shows sample unique patterns extracted for the new-thyroid and Hayes-Roth dataset.

Table 2. Extracted exclusive patterns for the minority classes of new-thyroid and hayes-roth.

Dataset	Features	Minority Class
new-thyroid	Thyroid stimulating, Thyroxin, T3resin, Triiodothyronine	Hypo
new-thyroid	Thyroxin, Triiodothyronine, T3resin	Hyper
hayes-Roth	Age, Educational Level, Marital Status	Case3

The extracted patterns are stored according to their level of exclusiveness with the target class. The confidence value is used to decide how strongly the patterns are related to the target class. The confidence value for rule is derived using Equation (2),

$$\text{Confidence } (R \Rightarrow C) = P(C|R) \qquad (2)$$

Here, *R* is the rule and *C* is the class.

If confidence value equals one then there is a tight bonding between the target class and patterns. From the extracted patterns, a pattern base is created for each class. It stores all patterns in order of their bonding with the target class.

Our proposed modified naïve Bayesian algorithm - MiNB uses this pattern base with class conditional probability to predict the class label. Equation (3) is used for the traditional naïve Bayesian classifier.

$$\text{C(X)} = (c \in C)\ P(c) \prod_{i=1}^{m} P(xi\ |c). \qquad (3)$$

Here, *X* is the instance to be classified, *C* is the list of classes, *xi* is the ith attribute of instance *X*, and c is the class.

The mathematic equation for the MiNB is defined as,

$$\text{C(X, P)} = argmax(c \in C)\ P(c) \prod_{i=1}^{m} P(xi\ |c)\ W \qquad (4)$$

Here, *P* is the pattern base and *W* is the membership weight derived using Equation (5). Other terms used here are the same as defined for Equation (3).

$$\text{W = MIR} * \sum_{r \in P} Z \mid \text{Z=1 if xi} \subset \text{r, otherwise Z=0} \qquad (5)$$

Here, *MIR* is the imbalanced ratio (#maj ÷ #min), xi is the “Att=val” pair, P is pattern base, and *Z* is a variable.

The traditional naïve Bayesian is the majority class-biased algorithm. The weight factor W is used to alleviate the biasing towards the majority class. The MiNB can give equal importance to the majority and minority classes by giving more weightage to the minority class through causal pattern discovery.

5. Evaluation Parameters

Table 3 shows the list of measures used for the performance evaluation. True Positive (TP) is used to give the number of correctly classified positive instances. True Negative (TN) provides the total number of negative instances that are correctly classified. False-Positive (FP) indicates the negative instances which are wrongly classified as positive instance and False Negative (FN) indicates the positive instances which wrongly classified as negative instances. We used minority class as a positive class and majority class as a negative class.

Table 3. List of measures used for performance evaluation.

Sr. No.	Name	Formula
1	Accuracy	(TP + TN) / TP+TN+FP+FN)
2	F-value	2*((precision*recall) / (precision + recall))
3	G-mean (Geometric Mean)	$\sqrt[n]{(\prod_{i=1}^{n} ACi)}$ Where, n=number of classes and ACi= accuracy of class Ci.

Accuracy gives the overall performance of the algorithm. F-value is used to give a balanced result from precision and recall. The G-mean value is the geometric mean of the class accuracy. As per [14], it is good to compare classifiers with more than one evaluation measures. We have used Accuracy, F-value, G-mean, and Area Under the Curve (AUC) for the performance evaluation. AUC is widely used and considered as the most accurate performance measure for the imbalanced data classification [4]. AUC is defined as an area under the receiver operating characteristics curve.

6. Implementation Methodology

We have used the weka library for the implementation of the MiNB. The performance of MiNB is compared with the data level sampling technique called SCUT [10] and a cost-sensitive algorithms.

Astha Herna *et al*. [10] stated that the SCUT algorithm performed well as compared to the SMOTE, Cluster Under-sampling Technique (CUT), and random under-sampling. Our experiments show that our proposed cost-sensitive MiNB algorithm outperforms the SCUT algorithm.

We have used ten-fold cross-validation for testing and training. Table 4 shows the performance summary of the SCUT and MiNB. The MiNB shows the best performance for the Hayes-Roth, new-thyroid, and Ecoli datasets.

Table 4. Results of MiNB and SCUT.

Dataset	SCUT [10]			MiNB		
	F-value	G-mean	ROC	F-value	G-mean	ROC
Hayes-Roth	0.693	0.704	0.886	**0.810**	**0.869**	**0.963**
New-Thyroid	0.954	0.951	0.998	**0.995**	**0.996**	**0.998**
Ecoli	0.887	0.936	0.974	**0.960**	**0.990**	**0.997**
Pageblocks	0.871	0.920	**0.982**	**0.975**	**0.991**	0.974
Yeast	0.595	0.753	0.912	**0.823**	**0.963**	**0.963**
Wine Quality	0.417	0.639	0.801	**0.630**	**0.885**	**0.821**

The MiNB has the best G-mean value for all six datasets. The wine-quality and yeast datasets are comparatively large datasets. The MiNB shows a good result for such large datasets as well. SCUT has better ROC values as compared to MiNB for the pageblocks dataset.

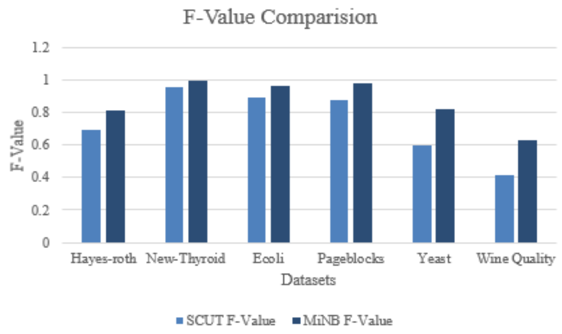

Figure 2. F-value comparison of SCUT and MiNB.

Figure 2 shows the f-value comparison graph. For all the dataset, the MiNB have a good f-value as compared to the SCUT. Figure 3 shows a comparison based on AUC value. The MiNB has the best AUC value for all the datasets except the Page blocks dataset.

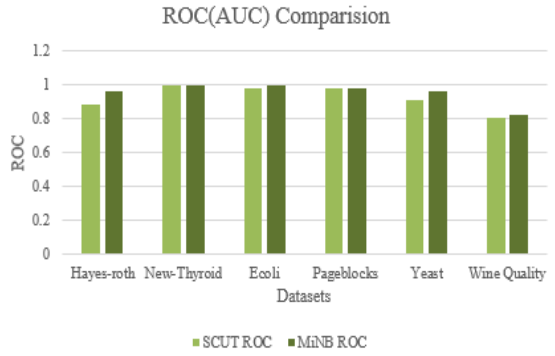

Figure 3. AUC Comparison of SCUT and MiNB.

Table 5 shows the class-specific performance of the MiNB for each dataset. We found that the MiNB shows good performance for the minority class. Earlier studies focused more on the majority class and the overall performance parameters such as accuracy and G-mean are also biased towards the majority class. However, the MiNB shows optimal performance for both the majority and minority classes of the dataset. Class-specific performance shown in Table 5 proves this unbiased performance of the MiNB. The overall performance of the MiNB doesn't dominate by the majority class performance. In unbalanced data learning, accurate prediction of the minority class is more important due to its high misclassification cost. If performance is evaluated by only traditional majority class-biased performance parameters then it is not possible to get the actual performance of the minority class.

Table 5. Class-specific performance of MiNB.

Dataset	**Class**	**ROC**	**TP Rate**	**Accuracy**	**Misclassification Rate**
Ecoli	cp	0.998	0.993	0.982	0.018
	im	0.997	0.974	0.982	0.018
	pp	0.996	0.865	0.976	0.024
	imU	0.999	0.942	0.988	0.012
	om	1.0	1.0	0.997	0.003
	omL	1.0	1.0	1.0	0
	imL	1.0	1.0	1.0	0
	imS	0.910	0.5	0.997	0.003
Hayes-roth	Class1	0.951	0.725	0.810	0.19
	Class2	0.954	0.784	0.810	0.19
	Class3	1.0	1.0	1.0	0
New-Thyroid	Normal	0.998	1.0	0.995	0.005
	Hyper	0.998	0.971	0.995	0.005
	Hypo	1.0	1.0	1.0	0
Page blocks	Type1	0.974	0.997	0.979	0.021
	Type2	0.988	0.878	0.990	0.01
	Type3	1.0	1.0	1.0	0
	Type4	0.983	0.375	0.990	0.01
	Type5	0.916	0.833	0.994	0.006
Yeast	MIT	0.969	0.79	0.944	0.05
	NUC	0.956	0.736	0.897	0.103
	CYT	0.953	0.889	0.859	0.141
	ME1	0.998	0.977	0.996	0.004
	ME2	0.974	0.803	0.991	0.009
	ME3	0.989	0.926	0.975	0.025
	EXC	0.996	0.914	0.997	0.003
	VAC	0.927	0.5	0.989	0.011
	POX	0.949	0.75	0.996	0.004
	ERL	1.0	1	1	0
Wine Quality	Type 3	0.98	0.966	0.998	0.002
	Type 4	0.941	0.8	0.986	0.014
	Type 5	0.86	0.694	0.782	0.218
	Type 6	0.759	0.587	0.668	0.332
	Type7	0.854	0.551	0.83	0.17
	Type 8	0.941	0.689	0.989	0.011
	Type 9	0.968	0.8	0.999	0.001

Table 5 also shows the rate of misclassification. The misclassification rate is very low for the majority and minority classes. This shows that the MiNB improves the performance of classification without lowering the performance of the minority class. Because of the reduction in misclassification rate the total misclassification cost also gets reduce.

Table 6 shows minority class performance comparison of the MiNB with AUC4.5, SC4.5 [15], and AECID [5]. We considered Ecoli, Yeast, Wine-quality, and Page-blocks datasets for this performance analysis.

The AUC4.5 uses AUC value for the selection of best splitting criteria. It selects the splitting criteria which maximize the AUC value.

Table 6. Performance comparison of MiNB with AUC4.5, CSC4.5, and SC4.5 for minority class.

Dataset	**MiNB**	**AUC4.5**	**AECID [5]**	**SC4.5**
Ecoli	**0.909**	0.857	0.807	0.714
Yeast	**0.829**	0.704	0.499	0.469
Wine quality	0.727	**0.734**	-	0.478
Page blocks	0.818	**0.925**	1.0	0.818
Average Performance Rate	**0.821**	0.805	0.768	0.620

The MiNB outperform the AUC4.5, SC4.5, and AECID for the Ecoli and Yeast datasets. For the Wine-quality and Page blocks dataset, the AUC4.5 shows the best result. As shown in Figure 4, the MiNB and the AUC4.5 show optimal results. However, in the case of MiNB, the average performance rate is high as compared to the other three cost-sensitive learning techniques.

For the comprehensive performance evaluation, the MiNB is compared with the ensemble technique called enhance Bagging (eBagging) proposed by Tuysuzoglu and Birant [26]. Table 7 shows the accuracy comparison of the proposed MiNB and eBagging.

When compared to the eBagging ensemble technique, the MiNB provides the more optimal accuracy. Accuracy is a generalized evaluation criterion. The MiNB maintains overall accuracy while improving minority class prediction rate, demonstrating its robustness in unbalanced data classification.

Table 7. Performance comparison of MiNB and eBagging.

Dataset	**MiNB**	**eBagging [26]**
Ecoli	**0.909**	0.879
Wine quality	0.727	**0.953**
Page blocks	**0.979**	0.967
Breast Cancer	**0.973**	0.734

7. Conclusions

In this research, a novel strong causal relationship based weighted naïve Bayesian classification model has been proposed for the unbalanced data learning. The proposed MiNB algorithm improves the accuracy of unbalanced data classification by improving the predictive performance of the minority class. The MiNB algorithm reduces overall misclassification cost for the domains where the minority class has more misclassification cost as compared to the majority class.

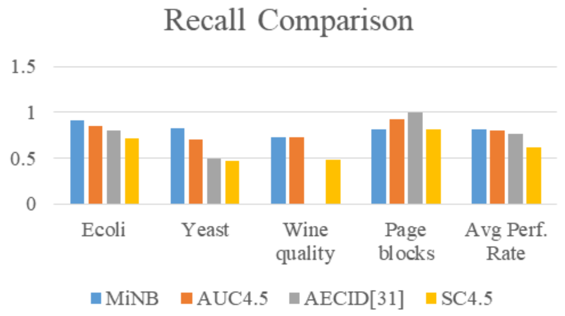

Figure 4. Recall Comparison of MiNB with AUC4.5, AECID, and SC4.5.

The MiNB compared with two techniques of unbalanced data learning – cost-sensitive learning and data sampling. The result of MiNB shown the improved performance as compared to the data balancing technique.

The MiNB has shown good performance when compared with the cost-sensitive approaches like AUC4.5 [15], SC4.5 [15], and AECID [5].

The MiNB correctly predicts the minority class with good accuracy even if the dataset is unbalanced. The unbiased predictive performance of cost-sensitive learning ultimately results in the reduction of misclassification costs. The main objective of this research is to propose an unbalanced learning approach which improves performance for minority class and thereby assist in the development of cost-effective machine learning tool for the domains where the minority class is more important as compared to the majority class.

References

[1] Al-Qerem A., Al-Naymat G., Alhasan M., and Al-Debei M., “Default Prediction Model: the Significant Role of Data Engineering in the Quality of Outcomes,” *The International Arab Journal of Information Technology*, vol. 17. no. 4A, pp. 635-44, 2020.

[2] Barot P. and Jethva H.,“Statistical Study to Prove Importance of Causal Relationship Extraction in Rare Class Classi fi Cation,” *in Processdings of The International Conference on Information and Communication Technology for Intelligent Systems*, Ahmedabad, pp. 416-425, 2017.

[3] Bashir K., Li T., and Yahaya M., “A Novel Feature Selection Method Based on Maximum Likelihood Logistic Regression for Imbalanced Learning in Software Defect Prediction,” *The International Arab Journal of Information Technology*, vol. 17, no. 5. pp. 721-730, 2020.

[4] Braytee A., liu W., and Kennedy P., “A Cost-Sensitive Learning Strategy for Feature Extraction from Imbalanced Data,” *in Processdings of 23rd International Conference on Neural Information Processing*, Kyoto, pp. 78-86, 2016.

[5] Chaabane I., Guermazi R., and Hammami M., “Enhancing Techniques for Learning Decision Trees from Imbalanced Data,”*Advances in Data Analysis and Classification*, vol. 14, no. 3, pp. 677-745, 2020.

[6] Chomboon K., Kerdprasop K., and Kerdprasop N., “Rare Class Discovery Techniques for Highly Imbalanced Data,” *in Proceedings of the International Multi Conference of Engineers and Computer Scientists*, Hong Kong, pp. 269-272, 2013.

[7] Chawla N., Bowyer K., Hall L., and Kegelmeyer W., “SMOTE: Synthetic Minority Over-Sampling Technique,” *The Journal of Artificial Intelligence Research*, vol. 16, pp. 321-57, 2002.

[8] Cieslak D. and Chawla N., “Learning Decision Trees for Unbalanced Data,” *in Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Antwerp, pp. 241-256, 2008.

[9] Garc´ıa S. and Herrera F., "Evolutionary Under-Sampling for Classification with Imbalanced Data Sets: Proposals and Taxonomy," *Evolutionary Computation*, vol. 17, no. 3, pp. 275-306, 2008.

[10] Herna L., Agrawal A., Viktor H., and Paquet E., "SCUT : Multi-Class Imbalanced Data Classification using SMOTE and SCUT : Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling," *in Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management-KDIR*, Lisbon, pp. 226-234, 2015.

[11] Jiang K., Lu J., and Xia K., "A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE," *Arabian Journal for Science and Engineering*, vol. 41, no. 8, pp. 3255-66, 2016.

[12] Kong G., Jiang L., and Li C., "Beyond Accuracy: Learning Selective Bayesian Classifiers with Minimal," *Pattern Recognition Letters*, vol. 80, pp. 165-71, 2016.

[13] Kotsiantis S., Kanellopoulos D., and Pintelas P., "Handling Imbalanced Datasets : A Review," *GESTS International Transactions on Computer Science and Engineering*, vol. 30, no. 1, pp. 25-36, 2006.

[14] Labatut V. and Cherifi H., "Accuracy Measures for the Comparison of Classifiers," *in Proceedings of 5th International Conference on Information Technology,* Amman, pp. 1-5, 2012.

[15] Lee J., "AUC4.5: AUC-Based C4.5 Decision Tree Algorithm for Imbalanced Data Classification," *IEEE Access*, vol. 7, pp.106034-106042, 2019.

[16] Muchlinski D., Siroky D., He J., and Kocher M., Comparing Random Forest with Logistic Regression for Predicting Class-Imbalanced Civil War Onset Data," *Political Analysis*, vol. 24, no. 1, pp. 87-103, 2016.

[17] Mujalli R., López G., and Garach L., "Bayes Classifiers for Imbalanced Traffic Accidents Datasets," *Accident Analysis and Prevention,* vol. 88, pp. 37-51, 2016.

[18] Patel H. and Thakur G., "Improved Fuzzy-Optimally Weighted Nearest Neighbor Strategy to Classify Imbalanced Data," *International Journal of Intelligent Engineering and Systems* vol. 10, no. 3, pp.156-162, 2016.

[19] Ratanamahatana C. and Gunopulos D., "Scaling up the Naive Bayesian Classifier: Using Decision Trees for Feature Selection," *in Proceedings of Work Data Clean Preprocessing (DCAP 2002), IEEE Intrenational Confrence Data Min*, pp. 613-623, 2002.

[20] Sáez J., Luengo J., Stefanowski J., and Herrera F., "SMOTE-IPF: Addressing the Noisy and Borderline Examples Problem in Imbalanced Classification By A Re-Sampling Method with Filtering," *Information Sciences*, vol. 291, pp. 184-203, 2015.

[21] Stefanowski J., *Dealing with Data Difficulty Factors While Learning from Imbalanced Data*, Springer, 2016.

[22] Sun Z., Song Q., Zhu X., Sun H., Xu B., and Zhou Y., "A Novel Ensemble Method for Classifying Imbalanced Data,"*Pattern Recognit,* vol. 48, no. 5, pp. 1623-37, 2015.

[23] Taheri S., Yearwood J., Mammadov M., and Seifollahi S., "Attribute Weighted Naive Bayes Classifier Using a Local Optimization," *Neural Computing and Applications*, vol. 24, pp. 995-1002, 2014.

[24] Triguero I., Del Río S., López V., Bacardit J., Benítez J., and Herrera F., "ROSEFW-RF: The Winner Algorithm for the ECBDL'14 big data Competition: An Extremely Imbalanced Big Data Bioinformatics Problem," *Knowledge-Based Systems*, vol. 87, pp. 69-79, 2015.

[25] Trisanto D., Rismawati N., Mulya M., and Kurniadi F., "Effectiveness Undersampling Method and Feature Reduction in Credit Card Fraud Detection," *International Journal of Intelligent Engineering and Systems*, vol. 13, no. 2, pp. 173-81, 2020.

[26] Tuysuzoglu G. and Birant D., "Enhanced Bagging (eBagging): A Novel Approach for Ensemble Learning," *The International Arab Journal of Information Technology*, vol. 17, no. 4, pp. 635-44, 2020

[27] Vluymans S., Triguero I., Cornelis C., and Saeys Y., "EPRENNID: An Evolutionary Prototype Reduction Based Ensemble for Nearest Neighbor Classification of Imbalanced Data, *Neurocomputing*, vol. 2016, pp. 596-610, 2016.

[28] Vural M. and Gok M., "Criminal Prediction Using Naive Bayes Theorym" *Neural Computing and Applications*, vol. 28, pp. 2581-2592, 2017.

[29] Wan C., "Test-Cost Sensitive Classification on Data with Missing Values in the Limited Time," *in Proceedings of the International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, Cardiff, pp. 501-510, 2010.

[30] Weiss G., McCarthy K., and Zabar B., "Cost-Sensitive Learning Vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?" *in Proceedings of the International Conference on Data Mining*, Las Vegas, pp. 1-7, 2007.

[31] Zaidi N., Cerquides J., Carman M., and Webb G., "Alleviating Naive Bayes Attribute Independence Assumption by Attribute Weighting,"*Journal of Machine Learning Research*, vol. 14 pp. 1947-1988, 2013.

[32] Zhang D., Ma J., Yi J., Niu X., and Xu X., "An Ensemble Method for Unbalanced Sentiment Classification,"*in Proceedings of the 11th*

International Conference on Natural Computation, Zhangjiajie, pp. 440-445, 2016.

Pratikkumar Barot received the B.E. degree From H.N.G.U and M.E. in computer engineering from Gujarat Technological University, India. He did his Ph.D. from Gujarat Technological University, India. His research interests include unbalanced data classification, machine learning, data mining, AI, data science and algorithm design.

Harikrishna Jethva currently works at the Head of Department, Department of Computer Engineering, Government Engineering College, Patan, Gujarat, India, His research interest in Machine Learning, Neural Network, Theory of Computation, Compiler Design, Soft Computing & Algorithms. He is Ph. D. Guide in Gujarat Technological University. In addition, he is a Board of Study member in many universities.