A Cluster-based Undersampling Technique for Multiclass Skewed Datasets

Rose Mary Mathew
Department of Computer Science, Karpagam Academy of Higher Education, India
rosem.mathew@gmail.com (corresponding author)

Ranganathan Gunasundari
Department of Computer Applications, Karpagam Academy of Higher Education, India
gunasoundar04@gmail.com

Received: 12 March 2023 | Revised: 29 March 2023 and 4 April 2023 | Accepted: 8 April 2023

 DOI: https://doi.org/10.48084/etasr.5844

ABSTRACT

Imbalanced data classification is a demanding issue in data mining and machine learning. Models that learn with imbalanced input generate feeble performance in the minority class. Resampling methods can handle this issue and balance the skewed dataset. Cluster-based Undersampling (CUS) and Near-Miss (NM) techniques are widely used in imbalanced learning. However, these methods suffer from some serious flaws. CUS averts the impact of the distance factor on instances over the majority class. Near-miss method discards the inter-class data within the majority of class elements. To overcome these flaws, this study has come up with an undersampling technique called Adaptive K-means Clustering Undersampling (AKCUS). The proposed technique blends the distance factor and clustering over the majority class. The performance of the proposed method was analyzed with the aid of an experimental study. Three multiminority datasets with different imbalance ratios were selected and the models were created using K-Nearest Neighbor (kNN), Decision Tree (DT), and Random Forest (RF) classifiers. The experimental results show that AKCUS can attain better efficacy than the benchmark methods over multiminority datasets with high imbalance ratios.

Keywords-K-means clustering; multiclass; resampling; skewed; undersampling

I. INTRODUCTION

Most of the real-world available data, e.g. for medical diagnosis, credit card transactions, fault detection, and activity recognition, are imbalanced. The term imbalance is a reflection of the difference in frequency of data distribution over the various classes in the dataset. The frequency of data present in one class is often higher than the frequency of the other classes. In terms of probability, the prior probabilities of various classes are different. For a binary class imbalanced problem, there is one minority class and one majority class. Training a model with this imbalanced data induces a bias towards the majority class. For a multi-class environment, imbalance has a bit different presentation. Data imbalance can occur in one of two ways: one majority class with numerous minority classes (multiminority case) and one minority class with numerous majority classes (multimajority case) [1].

Three things should be noted while working with imbalanced data. (a) There is a great difference in the frequency of the elements present in majority and minority classes. (b) Minority class data have a small probability of occurrence, so it is challenging to obtain new minority instances from the real world. (c) The classification result is much more influenced by the majority class instances, so the classification model is biased towards the majority data [2]. However, the interesting fact is that while the overall accuracy of the model is high, the minority classification is near to zero. This is corrected with the aid of resampling techniques.

Resampling techniques make the skewed dataset balanced. Resampling can be done in the form of oversampling, undersampling, or in hybrid mode. Oversampling techniques add new synthetic data to the minority classes for making the dataset balanced. Undersampling techniques eliminate data from the majority classes. Hybrid techniques are combinations of both to make the dataset balanced [3]. Oversampling technique expands the real data with synthetic samples and these synthetic data may overlap with the actual instances of other classes. Random oversampling, Synthetic Minority Oversampling Technique (SMOTE), and Adaptive Synthetic (ADASYN) are the most popular oversampling techniques. Undersampling techniques eliminate data from the majority class and this elimination will sometimes be reflected in the form of prominent data loss [4]. For balancing the skewed data, an appropriate resampling technique can be chosen from the available techniques.

The major contribution of this study is the proposal of an undersampling technique for multiminority dataset

classification. The proposed work is implemented with adaptive K-means clustering algorithm [5]. The proposed technique is named Adaptive K-means Clustering Undersampling (AKCUS). Another important factor to be considered is the imbalance ratio of the datasets which is used for model training. Another contribution of this work is the reduction of the ratio of skewness among various classes. The effectiveness of the proposed algorithm over 3 multiminority datasets with imbalance ratios of different levels is described.

II. RELATED WORKS

Several undersampling techniques are proposed in the literature. Most of them recognize the less informative occurrences, on account of the redundancy involved. Tomek links are pairs of instances which belong to different classes and are mutually nearest neighbors. These Tomek links from the dataset were eliminated in [6]. Edited Nearest-Neighbor (ENN) rule, removes those data points which belong to different classes from a majority set with K nearest neighbors [7]. Condensed Nearest Neighbor (CNN) is another undersampling technique which chooses a subset from the training set such that each data in the training set and its nearest neighbors in the subset belong to the same class. This technique is slower as it requires multiple passes over the training dataset [8]. Near-Miss (NM) method performs data undersampling in the majority class [9]. This is done by identifying their distance to every other data in the same class. Authors in [10] proposed the UnderBagging technique, in which in every iteration the dataset is undersampled randomly. Sorting of undersampled data depending on the corresponding weighted Euclidean distance from the minority data was proposed in [11]. Authors in [12] proposed a method which uses a collection of hardness measures for classification and they found that instance hardness is mainly caused by the overlapping of classes. Authors in [13] proposed a cluster-based approach of undersampling to balance the dataset and the classification was done with the aid of back propagation neural networks [13]. Authors in [14] proposed a hybrid algorithm which combines methods like EUS, AdaBoost, cost sensitive modification, and adaptive boundary decision strategy. EUSBoost [15] incorporates Random Undersampling (RUS) with Boosting algorithm and outperforms the existing proposals with the use of evolutionary undersampling approach. EasyEnsemble [16] is an ensemble of ensembles which trains a classifier for every new bag, and each bag is trained by AdaBoost. For multiclass skewed distributions, no specific methods are proposed.

In this study, a resampling mechanism for multiminority datasets is proposed and its effectiveness is measured.

III. PROPOSED WORK

Undersampling is the process of removing excess data from a skewed dataset in order to make it balanced. Majority class data are considered for clustering and in this work, adaptive k-means clustering is done over the majority class. The different clusters are identified, and data elimination was conducted over these clusters. The proposed method can be applied in real world multiminority datasets.

In cluster center-based undersampling, the primary stage is to split the skewed binary class dataset into training data and testing data. The training data are arranged based on the majority and minority sets. Undersampling is applied in order to eliminate the excess number of elements present in the majority set. In this method, the majority data are portioned into different bags and on each bag RUS is performed. Each undersampled bag is clubbed with the minority class for training the model. K clusters are formed, this K being the same with the number of elements present in the minority set (K=N). These K clusters are produced over M samples of the majority set. The entire majority set is replaced by cluster centers. The mean of data present in a cluster is treated as the cluster center. The nearest K neighbors of the cluster center are removed in order to obtain a balance among the majority class and the minority class [17].

The proposed Adaptive K-Means Clustering Undersampling (AKCUS) technique uses the Adaptive K-means algorithm. This algorithm calculates the cluster center effectively. The center points for multiclass data clusters can be identified by the algorithm. The majority class data are considered for clustering. This algorithm selects K elements from the set of the majority class. These selected elements are treated as seed. The properties defined by each element define the properties of the corresponding cluster. Euclidean distance is used to compute the distance [15]. The distance between each element and a cluster, the distance between two elements, and the distance between two clusters are computed. For two elements, P_1 and P_2, with n dimensions, the distance can be calculated as:

$$D = \sqrt{(P_{11} - P_{21})^2 + (P_{12} - P_{22})^2 + \ldots\ldots (P_{1n} - P_{2n})^2} \quad (1)$$

where $P_1 = (P_{11}, P_{12}, P_{13}, \ldots. P_{1n})$ and $P_2 = (P_{21}, P_{22}, P_{23}, \ldots. P_{2n})$.

In this algorithm, the distance between the clusters is identified and the value is stored in an array in the form of a triangular matrix. The shortest distance between the clusters D_{min} is recognized and the nearest clusters are also recognized. Cluster formation, corresponding centroid calculations, and farthest data elimination are mentioned in the algorithm:

Algorithm of the proposed AKCUS:

Step 1: Get the multiclass imbalanced dataset and identify the majority and minority classes.

Step 2: Compute the mean frequency of the minority class records and let it be stored in a variable called Freq.

Step 3: Identify the excess data of the majority class from the Freq data and let it be stored in a variable called excess.

Step 4: Take one majority class data for under-sampling.

Step 5: Apply adaptive K-means clustering algorithm to find the cluster points of the majority class.

1. Convert the data format from array to vector.
2. Select k elements from the majority class and assign these elements as seed of the cluster.

3. Compute the distance between each element and seed using the Euclidean distance formula.
4. Compute the threshold value of the shortest distance between clusters D_{min}.
5. Compare every distance with this threshold value and then update the cluster centroid.
   a. For any element P_i, which does not belong to any cluster, the distance of P_i from each cluster is computed. If the distance is 0 for a cluster, P_i is assigned to that cluster, and the algorithm proceeds to the next non-clustered element.
   b. Element P_i is assigned to the closest cluster when the distance between P_i and the cluster is less than distance D_{min}.
   c. If the value of D_{min} is below the distance of the element from the nearest cluster, the two closest clusters C_1 and C_2 are selected and are merged in C_1. Cluster C_2 is eradicated by eliminating all the instances belonging to it and its image is removed. After that, the new instance is added to this empty cluster for creating a new cluster.
6. The above steps are repeated based on the threshold value, and the centroid values are updated.

Step 5: Find the distance between each element and the corresponding centroid of the cluster.

Step 6: The elements with the largest distance are discarded from the dataset.

Step 7: Apply Step 6 for excess/k number of times in each cluster.

Step 8: Apply Step 4 to Step 7 for the other majority class data.

Step 9: Finally concatenate the minority class subset and reduced majority class subsets.

IV. EXPERIMENTAL STUDY

The proposed resampling technique in conjunction with various machine learning algorithms was used in this study. The input dataset was extracted from the KEEL data repository. Multiminority datasets with different imbalance ratios were chosen. The implementation was done in MATLAB. Figure 1 shows the different stages of the experiment. The different stages are:

- Selecting three multiminority datasets with different imbalance ratios from the KEEL repository. These datasets are of 3-class classification.
- Data pre-processing was done over the selected datasets for removing noisy data.
- The data were split in 75:25 training-testing ratio.
- Classifiers like K-Nearest Neighbors (K-NN), Decision Tree (DT), and Random Forest (RF) were applied, and the corresponding model performance was evaluated.
- Resampling techniques like AKCUS, RUS, ENN, and NM were applied to the training data.
- The resampled data with different techniques were used to create models with K-NN, DT, and RF and the model performance was evaluated.
- The performances of various models before and after applying various resampling techniques are analyzed.

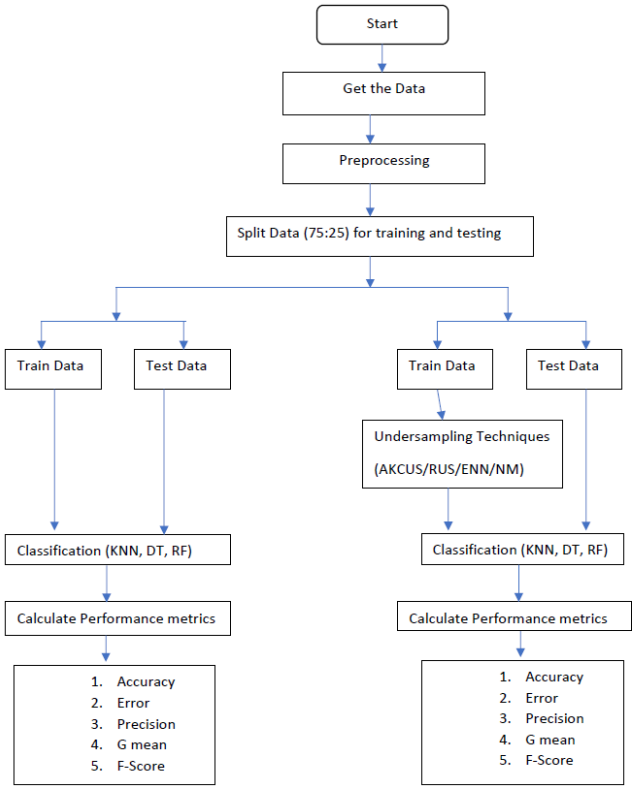

Fig. 1. Various stages in the study.

A. Data Used

A list of experiments was conducted to study the effectiveness of the proposed AKCUS algorithm. The empirical review was conducted on multiclass skewed datasets available in the KEEL data repository [18, 19]. Three multiminority datasets, namely New-Thyroid, Thyroid, and Wine were chosen. The information related to these datasets is given in Table I. The original data distribution plots of the datasets are shown in Figure 2.

B. Resampling Methods Used

In the data pre-processing stage, records were validated, and imbalanced datasets were transformed into balanced with the aid of resampling techniques. The proposed AKCUS was applied over the datasets and the classification results were compared with those of the existing undersampling techniques RUS, ENN, and NM.

C. Selected Classifiers

K-NN, DT, and RF were chosen for creating models in this experiment.

D. Performance Evaluation

The performance of different models that were built with the several resampling techniques and classifiers were assessed with the aid of the confusion matrix [20]. Confusion matrix denotes True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values. Using the confusion matrix, different metrics like accuracy, error, precision, F1-score, and G-mean of various models can be calculated.

TABLE I. DATASET INFORMATION

Dataset name	No of instances	Imbalance ratio	Attributes	Class label	No. of records
New-Thyroid	215	5	6	Normal	150
				Hyper	35
				Hypo	30
Thyroid	720	40.16	22	1	17
				2	33
				3	666
Wine	178	1.48	14	1	59
				2	71
				3	48

V. RESULTS AND DISCUSSION

The learning capability of different models can be identified with the help of performance metrics. In this study, all the datasets employ a data split of 75:25 ratio of model training and testing. The data distribution before and after applying the resampling methods is represented in Table II. Both RUS and NM bring similar numbers of elements after resampling. The selected datasets have multiple minority classes and a single majority class. In this experiment, the resampling occurs over the majority class.

TABLE II. DATA DISTRIBUTION BEFORE AND AFTER RESAMPLING

Dataset	Skewed data distribution			After undersampling								
				RUS/NM			ENN			AKCUS		
	c1	c2	c3	c1	c2	c3	c1	c2	c3	c1	c2	c3
New-Thyroid	147	35	37	20	20	20	113	20	18	25	25	25
Thyroid	17	37	666	11	11	11	11	3	458	12	12	12
Wine	57	71	38	29	29	29	40	38	29	30	30	30

To evaluate the effectiveness of AKCUS, a comparison was made between the performance metrics of different datasets. The results of a selected dataset with different balancing techniques as well as with different classifiers were considered. Initially, consider the New-Thyroid dataset and its performance metrics (Table III). It is evident that the model with RF classifier and AKCUS resampling technique over the New-Thyroid dataset achieved better performance than the other models. K-NN classifier works its best with the RUS sampled dataset. Classifiers DT and RF give promising results with AKCUS balanced datasets. Figure 3 represents the comparative chart of performance over various techniques applied over the New-Thyroid dataset.

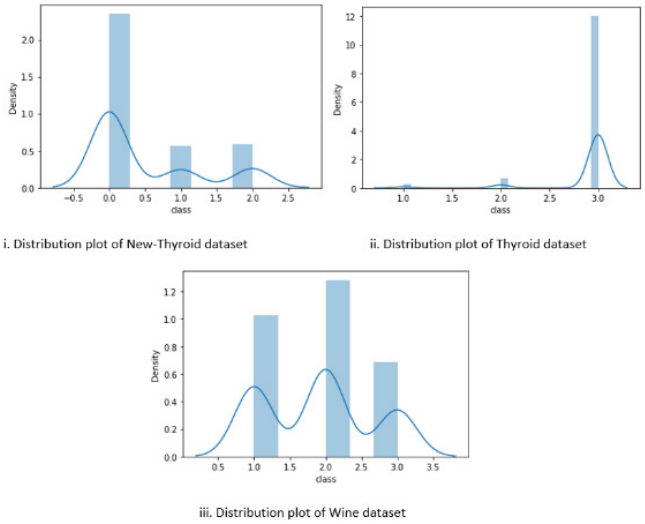

Fig. 2. Data distribution of the datasets.

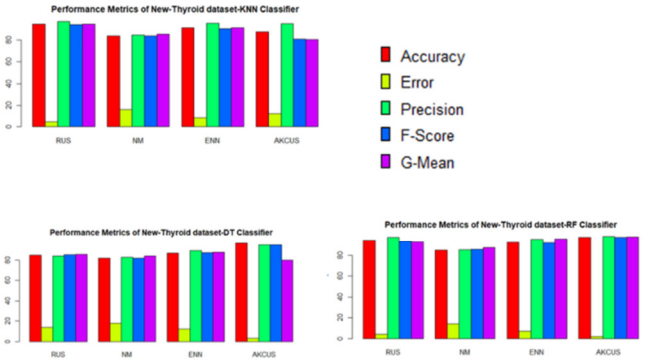

Fig. 3. Result comparison of various techniques over the New-Thyroid dataset.

TABLE III. PERFORMANCE METRICS OF NEW-THYROID DATASET

Resampling technique	Classifier	Accuracy	Error	Precision	F-Score	G-mean
Imbalanced set	K-NN	0.875	0.125	0.9477	0.8079	0.8051
	DT	0.9063	0.0938	0.9275	0.8725	0.8737
	RF	0.9531	0.0469	0.9783	0.9379	0.9289
RUS	K-NN	0.9454	0.0546	0.9687	0.9425	0.9387
	DT	0.8545	0.1455	0.8439	0.8549	0.8597
	RF	0.9531	0.0469	0.9783	0.9361	0.9305
NM	K-NN	0.8363	0.1637	0.8438	0.8369	0.8539
	DT	0.8181	0.1819	0.8303	0.8207	0.8414
	RF	0.8545	0.1455	0.8558	0.8567	0.8763
ENN	K-NN	0.9090	0.091	0.9509	0.9031	0.9109
	DT	0.8727	0.1273	0.893	0.8733	0.8763
	RF	0.9272	0.0728	0.9595	0.9213	0.9266
AKCUS	K-NN	0.875	0.125	0.9477	0.8079	0.8051
	DT	0.9688	0.0313	0.9589	0.9589	0.8051
	RF	**0.9818**	**0.0182**	**0.9888**	**0.9828**	**0.9832**

The thyroid dataset (Table IV) is another multiminority dataset. The model with DT classifier along with AKCUS resampling achieves better performance than the other models. KNN classifier performs its best with imbalanced data and AKCUS balanced dataset. DT and RF classifiers achieve their

best with AKCUS balanced dataset. Figure 4 shows the comparative chart of the performance over various techniques applied over the Thyroid dataset.

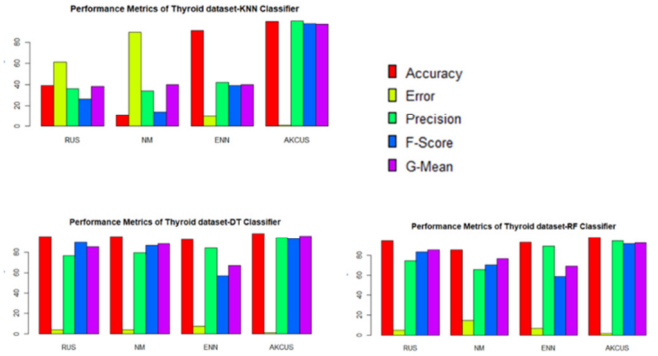

Fig. 4. Result comparison of the Thyroid dataset.

TABLE IV. PERFORMANCE METRICS OF THYROID DATASET

Resampling technique	Classifier	Accuracy	Error	Precision	F-Score	G-mean
Imbalanced set	K-NN	0.9954	0.0046	0.9983	0.9735	0.9673
	DT	0.9861	0.0139	0.949	0.9204	0.9295
	RF	0.9815	0.0185	0.8804	0.848	0.9004
RUS	K-NN	0.3944	0.6059	0.3623	0.2603	0.384
	DT	0.9611	0.0389	0.7679	0.9	0.8567
	RF	0.95	0.05	0.7476	0.8348	0.8567
NM	K-NN	0.1111	0.8889	0.3473	0.1443	0.4008
	DT	0.9611	0.0389	0.7976	0.8726	0.8867
	RF	0.8555	0.1445	0.6623	0.7052	0.7694
ENN	K-NN	0.9111	0.0889	0.4198	0.3927	0.4019
	DT	0.9277	0.0723	0.8455	0.569	0.6712
	RF	0.9333	0.0667	0.8957	0.5882	0.6922
AKCUS	K-NN	0.9954	0.0046	0.9983	0.9735	0.9673
	DT	0.9907	0.0093	0.9507	0.937	0.9673
	RF	0.9861	0.0139	0.949	0.9204	0.9295

The last multiminority dataset taken into consideration is the Wine dataset. Table V shows the metrics for the Wine dataset with different resampling methods and classifiers. The models with AKCUS did not perform better than the other models. KNN classifier performs its best with RUS sampled data. DT classifier achieved its best with NM and ENN. RF achieved its best with the imbalanced dataset, because the difference in the ratio of data among different classes is very low in Wine, so the proposed algorithm is not able to achieve better performance and this dataset can be considered as a special case. Figure 5 shows the comparative chart of the performance of various techniques over the Wine dataset.

The performance of the imbalanced dataset with different resampling techniques and classifiers are specified above. From the metric tables it was clear that AKCUS technique produces better results for all the datasets except the Wine dataset. The ratio of data imbalance was also considered while performing resampling. For the Wine dataset, the imbalance ratio is very low, so the balancing techniques were not much reflected. The imbalance ratio also acts as a factor while resampling measures are applied. Undersampling techniques sweep out some data that are available in the dataset and this will sometimes reflect badly on the data distribution. The proposed method sweeps out samples which are not very relevant for the learning process.

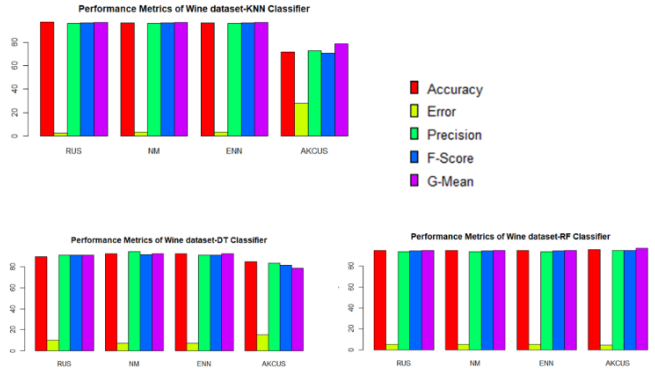

Fig. 5. Result comparison of the Wine dataset.

The Thyroid dataset, with a high imbalance ratio of 40.16, shows promising results with the proposed resampling technique. The New-Thyroid dataset with imbalance ratio of 5 shows better results with our proposed resampling technique. However, in the Wine dataset, with imbalance ratio of 1.48, the proposed method did not show much success.

TABLE V. PERFORMANCE METRICS OF WINE DATASET

Resampling technique	Classifier	Accuracy	Error	Precision	F-Score	G-mean
Imbalanced Set	K-NN	0.7042	0.2958	0.6951	0.687	0.7733
	DT	0.8592	0.1408	0.8478	0.8314	0.8756
	RF	0.97	0.03	0.96	0.947	0.95
RUS	K-NN	**0.9761**	**0.0239**	**0.9666**	**0.9717**	**0.9728**
	DT	0.9047	0.0953	0.9166	0.9166	0.9166
	RF	0.95	0.05	0.939	0.947	0.9494
NearMiss	K-NN	0.97	0.03	0.966	0.97	0.9724
	DT	0.9285	0.0715	0.9473	0.9196	0.9265
	RF	0.95	0.05	0.939	0.947	0.9494
ENN	K-NN	0.97	0.03	0.966	0.97	0.9724
	DT	0.9285	0.0715	0.9166	0.9178	0.9269
	RF	0.95	0.05	0.939	0.947	0.9494
AKCUS	K-NN	0.7183	0.2817	0.7313	0.7054	0.7928
	DT	0.8451	0.1549	0.8397	0.8205	0.7928
	RF	0.9577	0.0423	0.95	0.9534	0.9722

VI. CONCLUSION

Machine learning models for classification created with imbalanced data have a bias towards the majority class, affecting the classification process. In this study, a new undersampling technique named Adaptive K-means Clustering Undersampling (AKCUS) is proposed for multiclass skewed data classification. AKCUS performs well with multiminority cases of imbalanced data. As undersampling removes elements from the actual dataset, sometimes this removal will affect the loss of prominent data. The proposed AKCUS algorithm is concerned about the prominent elements in multiple classes and its removes data which are not very relevant to the classification process. In this work, the imbalance ratio was

considered as an important factor. The proposed algorithm works well with multiminority datasets that have high imbalance ratio. If the imbalance ratio is low, the result is not much affected by the resampling procedure. So, the proposed algorithm produces promising results for datasets with high imbalance ratio. This work can be extended to tackle the bias over big data models.

REFERENCES

[1] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, May 2017, https://doi.org/10.1016/j.eswa.2016.12.035.

[2] S. Tahzeeb and S. Hasan, "A Neural Network-Based Multi-Label Classifier for Protein Function Prediction," *Engineering, Technology & Applied Science Research*, vol. 12, no. 1, pp. 7974–7981, Feb. 2022, https://doi.org/10.48084/etasr.4597.

[3] W. M. S. Yafooz, E. A. Hizam, and W. A. Alromema, "Arabic Sentiment Analysis on Chewing Khat Leaves using Machine Learning and Ensemble Methods," *Engineering, Technology & Applied Science Research*, vol. 11, no. 2, pp. 6845–6848, Apr. 2021, https://doi.org/10.48084/etasr.4026.

[4] F. Belloum, L. Houichi, and M. Kherouf, "The Performance of Spectral Clustering Algorithms on Water Distribution Networks: Further Evidence," *Engineering, Technology & Applied Science Research*, vol. 12, no. 4, pp. 9056–9062, Aug. 2022, https://doi.org/10.48084/etasr.5116.

[5] S. Bhatia, "Adaptive K-Means Clustering," in *Proceedings of the Seventeenth International Florida Artificial Intelligence Research Society Conference*, Miami Beach, FL, USA, Jan. 2004, https://doi.org/10.13140/2.1.4197.9845.

[6] I. Tomek, "Two Modifications of CNN," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-6, no. 11, pp. 769–772, Aug. 1976, https://doi.org/10.1109/TSMC.1976.4309452.

[7] D. L. Wilson, "Asymptotic Properties of Nearest Neighbor Rules Using Edited Data," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. SMC-2, no. 3, pp. 408–421, Jul. 1972, https://doi.org/10.1109/TSMC.1972.4309137.

[8] P. Hart, "The condensed nearest neighbor rule (Corresp.)," *IEEE Transactions on Information Theory*, vol. 14, no. 3, pp. 515–516, Feb. 1968, https://doi.org/10.1109/TIT.1968.1054155.

[9] J. Zhang and I. Mani, "kNN approach to unbalanced data distributions: a case study involving information extraction," presented at the ICML, Washington DC, USA, 2003.

[10] R. Barandela, R. M. Valdovinos, and J. S. Sánchez, "New Applications of Ensembles of Classifiers," *Pattern Analysis & Applications*, vol. 6, no. 3, pp. 245–256, Dec. 2003, https://doi.org/10.1007/s10044-003-0192-z.

[11] A. Anand, G. Pugalenthi, G. B. Fogel, and P. N. Suganthan, "An approach for classification of highly imbalanced data using weighting and undersampling," *Amino Acids*, vol. 39, no. 5, pp. 1385–1391, Nov. 2010, https://doi.org/10.1007/s00726-010-0595-2.

[12] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Machine Learning*, vol. 95, no. 2, pp. 225–256, May 2014, https://doi.org/10.1007/s10994-013-5422-z.

[13] W. Liu, H. Zhang, Z. Ding, Q. Liu, and C. Zhu, "A comprehensive active learning method for multiclass imbalanced data streams with concept drift," *Knowledge-Based Systems*, vol. 215, Mar. 2021, Art. no. 106778, https://doi.org/10.1016/j.knosys.2021.106778.

[14] W. Lu, Z. Li, and J. Chu, "Adaptive Ensemble Undersampling-Boost: A novel learning framework for imbalanced data," *Journal of Systems and Software*, vol. 132, pp. 272–282, Oct. 2017, https://doi.org/10.1016/j.jss.2017.07.006.

[15] M. Galar, A. Fernández, E. Barrenechea, and F. Herrera, "EUSBoost: Enhancing ensembles for highly imbalanced data-sets by evolutionary undersampling," *Pattern Recognition*, vol. 46, no. 12, pp. 3460–3471, Dec. 2013, https://doi.org/10.1016/j.patcog.2013.05.006.

[16] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 39, no. 2, pp. 539–550, Apr. 2009, https://doi.org/10.1109/TSMCB.2008.2007853.

[17] S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," *Expert Systems with Applications*, vol. 36, no. 3, Part 1, pp. 5718–5727, Apr. 2009, https://doi.org/10.1016/j.eswa.2008.06.108.

[18] J. Alcala-Fdez *et al.*, "KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework," *Journal of Multiple-Valued Logic and Soft Computing*, vol. 17, no. 2, pp. 255–287, Jan. 2010.

[19] J. Alcalá-Fdez *et al.*, "KEEL: a software tool to assess evolutionary algorithms for data mining problems," *Soft Computing*, vol. 13, no. 3, pp. 307–318, Feb. 2009, https://doi.org/10.1007/s00500-008-0323-y.

[20] M. O. Ojo and A. Zahid, "Improving Deep Learning Classifiers Performance via Preprocessing and Class Imbalance Approaches in a Plant Disease Detection Pipeline," *Agronomy*, vol. 13, no. 3, Mar. 2023, Art. no. 887, https://doi.org/10.3390/agronomy13030887.

AUTHORS PROFILE

Rose Mary Mathew is a Research Scholar in the Karpagam Academy of Higher Education, Coimbatore. Currently she is working as an Assistant Professor (Special Grade) in the Department of Computer Applications, Federal Institute of Science and Technology, Angamaly. She has more than ten years of teaching experience. She obtained her Master of Computer Applications degree from Mahatma Gandhi University, Kottayam in 2009 and MBA from Bharathiyar University, Coimbatore in 2018. Her areas of specialization is Machine Learning, Deep Learning, and Artificial Intelligence. She has participated and presented several papers in National and International seminars and conferences.

Ranganathan Gunasundari is presently working as a Professor in the Department of Computer Applications, Karpagam Academy of Higher Education, Coimbatore. She has more than fifteen years of teaching experience. She has participated and presented several papers in National and International conferences. Her research interests include Data Mining, Cryptography and Network Security.