INDIAN JOURNAL OF SCIENCE AND TECHNOLOGY

RESEARCH ARTICLE

KNN Based Under Sampling: A Cognitive Centred Solution for Imbalanced Dataset Problem in Anaphora Resolution

K Arul Deepa[1]*, **Shanmuga Priya[2]**, **P Velvizhy[2]**

1 Dept. of IST, College of Engineering Guindy, Anna University, Chennai
2 Dept. of CSE, College of Engineering Guindy, Anna University, Chennai

Check for updates

OPEN ACCESS

Received: 20-06-2023

Accepted: 28-06-2023

Published: 08-08-2023

Citation: Deepa KA, Priya S, Velvizhy P (2023) KNN Based Under Sampling: A Cognitive Centred Solution for Imbalanced Dataset Problem in Anaphora Resolution. Indian Journal of Science and Technology 16(30): 2317-2324. https://doi.org/10.17485/IJST/v16i30.1523

* **Corresponding author**.

aruldeepa@auist.net

Funding: None

Competing Interests: None

Published By Indian Society for Education and Environment (iSee)

ISSN
Print: 0974-6846
Electronic: 0974-5645

Abstract

Background: Like many other real world applications, the machine learning system of anaphora resolution also struggles with skewed data. The problem of imbalanced classes occurs with classification task where there a huge difference exists in the number of instances among the involved classes.**Objectives:** The proposed framework intends to remove the imbalance first between positive and negative class instances before classifying them by KBUS that makes use of cognitive knowledge about the language and analysis is done at attribute level. **Method:** Nine pruning rules are crafted by KBUS(KNN Based Under Sampling) for TDIL dataset. **Findings:** During experimentation, number of positive instances are increased from 5.32% to 43.95%, whereas the number of negative instances are decreased from 94.68% to 56.05%. Loss ratio of positive and negative instance is 1:112. Finally the pruned dataset is classified by a list of classifiers namelyNaïve Bayes, SVM, Random forest, decision tree and k-NN. **Novelty:** Classifier results are discussed in two perspectives: Firstly the number of input instances and secondly the performance improvement achieved after pruning. It is adduced that pruning shows a remarkable improvement for all the classifiers. The proposed system produced an encouraging result as 78% of f-measure for k-NN and 77% for decision tree. Performance is presented in a comparative manner before and after pruning and the improvement of f-measureranges from 13% (k-NN) to 41% (Random Forest). Thus this work has come up with a machine learning model to resolve Tamil anaphoric situations effectively in an imbalanced classification environment.

Keywords: Imbalanced Dataset; Classification; Anaphora Resolution Machine Learning; Pronominal Reference

1 Introduction

The problem of imbalanced classes occur with the classification task in which there prevails a huge difference in the number of instances, among the involved classes. Like any other real world problems, such skewed dataset also reduces the machine learning capability of anaphora resolution system as well[1]. For any Natural Language Understanding tasks, resolving of pronominal references known as anaphora resolution is unavoidable. Anaphora is the pronominal reference to an entity that has been previously introduced into the discourse[2].

The existing work of Machine Learning System of Tamil Anaphora Resolution(ML-STAR) was an approach which attempted to resolve all the pronominal references in the given Tamil Text. The proposed approach is named as k-NN Based Under Sampling (KBUS) based on attribute level instance analysis, to prune the negative instances. Usually the under sampling approaches randomly remove majority class instances which sometimes result in the loss of potential instances of that class. Instant of random under sampling, the proposed KBUS provides justification for the removal of negative instances by forming 9 levels of pruning rules by attribute level cognitive analysis. Major assignment in the proposed work is to prune more negative instances with minimum lose in positive class instances. As the class instances are the pairs of anaphora and antecedent, the proposed system is named as Pair-Pruned System of Tamil Anaphora Resolution (PP-STAR). After inclusion of the pair pruning module, the STAR has shown a remarkable improvement in performance.

2 Methodology

There exist various methods to address this class imbalance problem. The two most common ways of addressing this issue are over sampling[3,4] and under sampling[5,6].

Let T be the training set with m attributes and n instances. If X is the set of all attributes (a1, … an) and Y is the set of all classes (1, …, c), T can be defined as, T = {xi, yj}, where xi is an instance in the set of attributes xi ∈ X and yi is an instance in the set of classes $yj \in Y$ Then T is said to be imbalanced if there is a subset with positive instances P ⊂ X and a subset of negative instances N ⊂ X and $|P| < |N|$.

Here in PP-STAR, a pruning module is incorporated to convert the imbalanced feature space to balanced feature space for better classification and is depicted in Figure 1 . As each instance of the imbalanced data set is a pair of anaphor and antecedent, the system is called as pair pruned STAR.

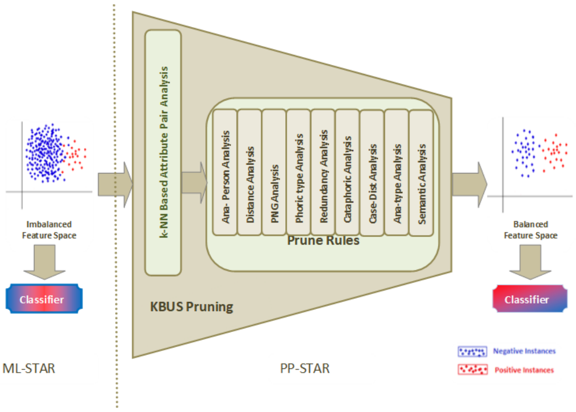

Fig 1. Design of PP-STAR

Random under-sampling[7] is an approach to handle the imbalanced dataset that balances the dataset by removing some randomly selected negative instances. But the major drawback of this is that it may discard some negative instances that are potentially helpful for the decision making of the classifier[8]. The proposed PP-STAR used the approach of under-sampling. But instead of random sampling this work justifies the pruning module by incorporating intelligence into the negative instance removal phase. This intelligence is obtained by analysing the characteristics of the feature space intrinsically through scatter plot of nearest neighbor analysis done between pair wise combination of features. The pruning module of PP-STAR is named as k-NN Based Under Sampling (KBUS). In all previous works, k-NN under-sampling[9] is done on the class level analysis, which considers the entire feature set. But the proposed PP-STAR establishes k-NN to the lower level analysis, which carries out the analysis between combinations of features. If a_i and a_j are any two attributes, scatter plot can be drawn between these two attributes to analyse the outlier instances with the minimum mislay in positive instances during under-sampling. Rules are derived through scatter plot analysis for the justification of pruning. The number of rules immediately relies on the feature value distribution in the feature space. For the considered dataset, nine rules are crafted for pruning.

2.1 KBUS Pruning

To achieve better learning experience there should be even distribution of samples for each class involved[10]. There is a remarkable mismatch obtained among positive and negative instances in all the datasets. After a deep analysis of the instances major root causes of the divergences are identified. To reconcile this discrepancy, in KBUS rules are framed by analysing the instances through scatter plots drawn between the various pairs of features. Scatter plots are constructed using Orange tool. From the useful plots 9 rules are framed as below. The number of rules is extendable depending upon the feature value distribution in the feature space. For the considered dataset the proposed work derived 9 rules which are explained below:

2.1.1 Rule-1: Pruning instances with first & second person pronouns

Due to its less frequency the pairs with first and second person pronouns are removed. At this juncture the exigency is to improve the classification accuracy, for this particular dataset, with due consideration given to third person pronouns. This step can also be avoided to resolve first and second person pronouns.

2.1.2 Rule-2: Tightening the distance

As the plot in Figure 3 says distance between anaphor and antecedent (number of lines in between) for positive pairs are 0 to 3. -4 to -1 and 4 to 5 line distances are occurred only for negative pairs. Therefore, sentence distance is reduced from (-4 to +5) to (0 to 3).

2.1.3 Rule – 3: Pruning PNG non-agreeing pairs

Tamil nouns and pronouns are sensitive to person, number and gender agreements. Thus PNG (Person, Number, Gender)- non agreeable pairs are removed. If the anaphor and antecedent pair do not agree even with any one of person, number or gender then such pair's row can be removed.

2.1.4 Rule – 4: Pruning exophora instances

Exophora is a kind of reference case where there would not be any direct antecedent. As this is another research area to be focused all the irresolvable pairs are removed (Phorics of the type exophora).

2.1.5 Rule – 5: Eliminating redundancy

Duplicates (in negative pairs) are removed, as the focus of pruning is to strengthen the positive class. All the repeated combinations are removed. There are some pairs that are generated repeatedly which in turn lead to more number of negative pairs.

Rule – 6: Removing cataphoric instances

It is observed that cataphorics are happening only with nominative antecedents. Therefore, all the other antecedent cases were removed. Thus for cataphora, only the pairs with nominative antecedents are considered.

2.1.7 Rule – 7: Pruning by distance - case analysis

With distance 2 and 3, only the pairs with antecedents with nominative case markers are considered.

2.1.8 Rule – 8: Pruning by distance – pronoun type analysis

With distance 3, pairs with reflexive pronouns are removed as they had never occurred with this distance 3

2.1.9 Rule – 9: Distance, semantic type analysis

With distance 2 and 3, only pairs of antecedent semantic type person & location are considered.

And the order of rules is also interchangeable. The projection evaluation method for scatter plot construction is k-Nearest Neighbour with 10-fold cross validation method. Table 1 presents the summary of prune rules derived for the present scenario.

Table 1. Summary of prune rules

Rule Number	Prune Rule	Considered pair of features	Effect
R0	Before Pruning	-	Imbalanced feature space
R1	**Rule 1**	<ana-person, class_label>	Pruning instances with I & II person pronouns
R2	**Rule 2**	<Dist, class_label>	Tightening the Distance as 0 to 3
R3	**Rule 3**	<PNG_agree, class_label>	Pruning PNG non-agreeing pairs
R4	**Rule 4**	<phoric_type, Result>	Pruning Exophora Instances
R5	**Rule 5**	Full feature vector	Eliminating Redundancy
R6	**Rule 6**	<Ant_case, phoric_type>	Removing Cataphoric Instances
R7	**Rule 7**	<Ant_case, dist>	Pruning instances with distance 2 / 3 and case is not nominative
R8	**Rule 8**	<ana_type, dist>	Pruning instances with reflexive pronouns and distance is 3
R9	**Rule 9**	<Ant_Stype, dist>	Pruning instances with distance 2 / 3 and semantic type is not equal to the person / location

3 Resultand Discussion

3.1 Results of KBUS

Experiments were done on the same setup of classifiers with the orange tool for both ML-STAR (i.e., the existing system called Machine Leaning Based System of Tamil Anaphora Resolution) and PP-STAR (i.e., the proposed system called Pair Pruned System of Tamil Anaphora Resolutionafter pruning). For explanatory purpose reduction of the ratio between YES class and NO class pair counts of TDIL training samples are given in Table 2. And the percentage of pair counts of each positive and negative pairs are explicitly presented in as chart in Figure 2.

Table 2. Performance of under-sampling – reduction at each rule

Target Class	Result	R0 (Before Pruning)	Prune Rules								
	# / %		R1	R2	R3	R4	R5	R6	R7	R8	R9
Yes	#	700	660	600	600	600	600	600	600	600	596*
	%	5.32	5.48	11.24	22.29	30.12	32.05	33.11	37.22	38.07	43.95
No	#	12452	11380	4740	2092	1392	1272	1212	1012	976	760*
	%	94.68	94.52	88.76	77.71	69.88	67.95	66.89	62.78	61.93	56.05
Total	#	13152	12040	5340	2692	1992	1872	1812	1612	1576	1356

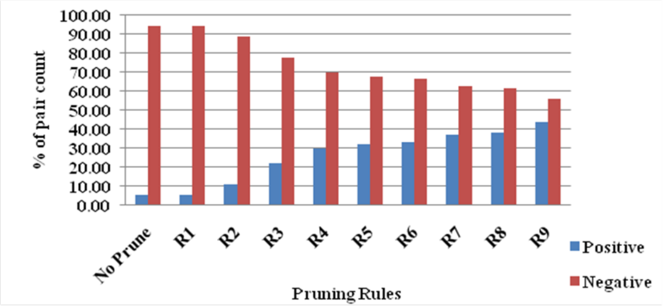

Fig 2. Performance of under-sampling

List of pruning rules is applied over the feature vector to reduce the difference between positive and negative pairs before it is given to classifiers[11]. During experimentation number of positive instances is increased from 5.32% to 43.95%, whereas the number of negative instances is decreased from 94.68% to 56.05%. In the proposed method the loss ratio of positive and negative instances is 1:112. After passing through 9 rules of pruning the dataset is made balanced between the positive and negative instance, which create a better learning experience for the classifiers.

3.2 Results of Classifiers

Performance of the classifiers is evaluated with the same metrics as that of ML-STAR: that is Precision (P), Recall (R) and f-measure (F). The classifier results are discussed in two perspectives: first is analyzing the performance with varying number of instances with k-NN and Decision tree classifiers[12,13]. Second is obtaining the classifier performance at various pruning rules from no prune (R0) to rule 9 (R9). These results are generated for the pruned feature space. Three versions of the dataset is prepared with varying sizes as DS1 with 339, DS2 with 635 and DS3 with1356 instances. Metric values are obtained for four validation methods including Cross-Validation (10-folds) (CV), Leave-One-Out (LOO), Random Sampling (10 repeats) with 70% propagation of training instances (RS), Test on Train Data (TTD).

By the attained results it is observed that classifier performance is directly proportional to the number of instances involved in the classification. Evaluation chart of classifier performances is presented in Figures 3 and 4. The results according to these charts clearly demonstrate that the learning algorithms performed much better with more training data. The highest f-measure is obtained with DS3 (TTD validation) as 78% for decision tree classifier and 85% for k-NN classifier.

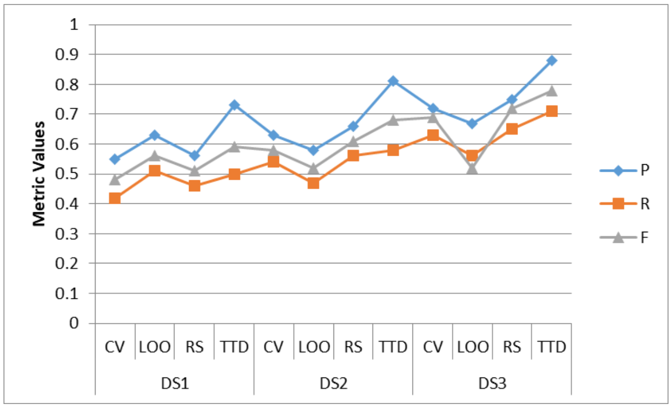

Fig 3. Performance of decision tree classifier

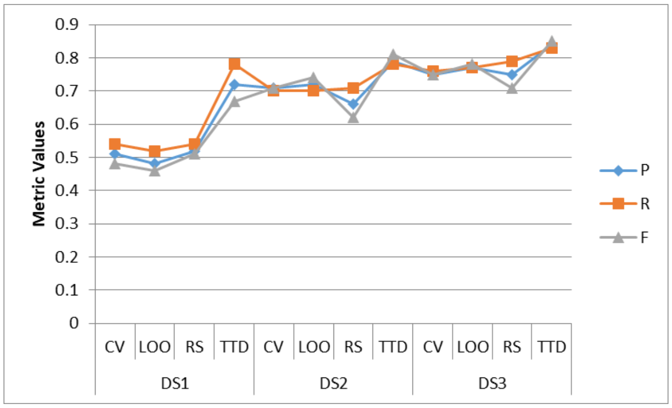

Fig 4. Performance of k-NN classifier

Except to the leave-one-out (LOO) validation method all the others are showing improvement with increase in the number of instances in the dataset. This is isdue to twinning in the data set; where there are some exactly nearby identical samples present. Recall analysis of k-NN is better with test on train data validation method.

Here we analyze the classifier performance after each prune rule. The precision, recall and f-measure values of various classifiers at various pruning levels are presented in Table 3. Here R0 denotes the results obtained before pruning. These results were obtained using 10 fold cross validation method[14]. Obtained results justify the importance of pruning rules. As anticipated, there is a fair increase in recall values.

Table 3. Precision, recall and f-measure at each prune rule

Classifiers	R0 (Before Pruning)	R1	R2	R3	R4	R5	R6	R7	R8	R9
Precision at each prune rule										
Naive Bayes	0.47	0.47	0.51	0.55	0.61	0.62	0.63	0.64	0.66	0.66
Random Forest	0.80	0.86	0.78	0.81	0.80	0.84	0.83	0.84	0.85	0.85
SVM	0.70	0.76	0.74	0.74	0.68	0.73	0.68	0.68	0.70	0.70
Classification Tree	0.73	0.74	0.73	0.72	0.77	0.81	0.80	0.82	0.84	0.84
kNN	0.72	0.74	0.73	0.72	0.79	0.79	0.78	0.79	0.79	0.79
Recall at each prune rule										
Naive Bayes	0.15	0.24	0.53	0.51	0.55	0.55	0.56	0.57	0.57	0.57
Random Forest	0.16	0.21	0.25	0.33	0.48	0.55	0.58	0.57	0.57	0.57
SVM	0.22	0.17	0.34	0.34	0.34	0.36	0.50	0.50	0.44	0.44
Classification Tree	0.38	0.42	0.45	0.45	0.59	0.68	0.66	0.75	0.71	0.71
kNN	0.60	0.61	0.65	0.66	0.70	0.72	0.80	0.73	0.77	0.77
f-Measure at each prune rule										
Naive Bayes	0.23	0.32	0.52	0.53	0.58	0.58	0.59	0.60	0.61	0.61
Random Forest	0.27	0.34	0.38	0.47	0.60	0.66	0.68	0.68	0.68	0.68
SVM	0.33	0.28	0.47	0.47	0.45	0.48	0.58	0.58	0.54	0.54
Classification Tree	0.50	0.54	0.56	0.55	0.67	0.74	0.72	0.78	0.77	0.77
kNN	0.65	0.67	0.69	0.69	0.74	0.75	0.79	0.76	0.78	0.78

3.3 Performance Analysis of Existing vs Proposed System

Notable improvement in f-measure is obtained with rules R1, R2, R3 and R4. Though there was 5% of negative instances pruned by R9, it doesn't show any impact in the f-measure of any of these classifiers and right now the importance of Rule 9 could not be evaluated. The observation is only for the considered dataset and the classifier setups. Table 4 presents the STAR performance in comparative manner of before and after pruning. i.e., performance comparison of ML-STAR (the existing system)and PP-STAR(the proposed system). The performance measure of each may better be analyzed with the chart given in Figure 5. Range of increase in f-measure values are highlighted by '*'. Maximum improvement is attained with Random forest. Though the least value of improvement is with k-NN as 13%, it is also a remarkable improvement. Thus pruning phase plays a vital role in all the classifiers irrelevant of the dataset or learning setup.

Table 4. STAR performance – ML-STAR (Existing System) Vs PP-STAR(Proposed System)

Classifier	Precision %			Recall %			f-Measure %		
	ML-STAR	PP-STAR	Range of increase	ML-STAR	PP-STAR	Range of increase	ML-STAR	PP-STAR	Range of increase
Naive Bayes	47	66	19	15	57	42	23	61	38**
Random Forest	80	85	5	16	57	41	27	68	41*
SVM	70	70	0	22	44	22	33	54	21
Decision Tree	73	84	11	38	71	33	50	77	27***

Continued on next page

Table 4 continued

k-NN	72	79	7	60	77	17	65	78	13

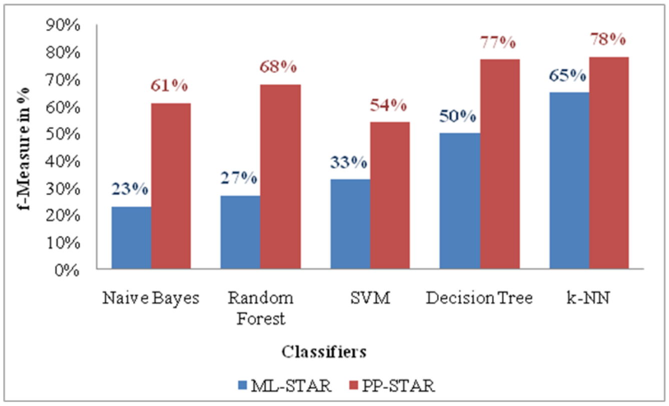

Fig 5. Improvement range off-measure (ML-STAR- Existing Vs PP-STAR-Proposed)

4 Conclusion

The problem of class imbalance occurs in a classifier system when there is a drastic difference in the number of instances among the involved classes. This leads to deterioration in classifier performance. In ML-STAR the ratio between the number of positive and negative instances was 1:19. In the proposed PP-STAR, an approach named KBUS is developed based on k-NN under-sampling at the attribute level for pruning the negative instances with minimum lose in positive instances. Novelty of this proposal is the application of cognitive level knowledge of domain to extract pruning rules that help in under sampling.9 levels of pruning rule are formed based on the scatter plot analysis done between various pairs of features. Depending on the classification environment and nature of the dataset, the number and order of prune rules are subject to vary. Number of positive instances was increased from 5.32%(in existing system) to 43.95%(in proposed system), whereas the number of negative instances was decreased from 94.68%(in existing system) to 56.05%(in proposed system). In the proposed method the loss ratio of positive and negative instances is 1:112. After equating the number of instances of positive and negative classes, they are passed to the classifiers namely, Naïve Bayes, SVM, Random forest, Decision Tree and k-NN. Initially the system was experimented with sample data of varying sizes for k-NN and decision tree and obtained f-measure as 85% and 78% respectively. From this result it is observed that classifiers performed better with increase in number of instances. Then the results of PP-STAR are compared with ML-STAR and the influence of KBUS pruning is justified. All the classifiers showed notable improvement in f-measure ranges from 13%(in existing system) to 41%(in proposed system). Thus an enhanced version of STAR is developed as PP-STAR by including the pair pruning module.Akin researches are emerging as the imperative and inevitable research applications in Tamil computing nowadays. This research has thrown light on complexities on NLU tasks like question answering, summarization, sentiment analysis, etc., and deliberates on their significances and the methodology to imply considerable impact on them. This work may be experimented with varying other tasks of NLP and even with other domains like e-Commerce, Intrusion Detection System, Cyber Crime Analysis etc., as future work.

References

1) Malik KEF. New Hybrid Data Preprocessing Technique for Highly Imbalanced Dataset. *Computing and Informatics*. 2022;41:981–1001. Available from: https://doi.org/10.31577/cai20224981.
2) Ezzini S, Abualhaija S, Arora C, Sabetzadeh M. Automated handling of anaphoric ambiguity in requirements. In: Proceedings of the 44th International Conference on Software Engineering. ACM. 2022;p. 187–199. Available from: https://doi.org/10.1145/3510003.3510157.
3) Kunakorntum I, Hinthong W, Phunchongharn P. A Synthetic Minority Based on Probabilistic Distribution (SyMProD) Oversampling for Imbalanced Datasets. *IEEE Access*. 2020;8:114692–114704. Available from: https://doi.org/10.1109/ACCESS.2020.3003346.
4) Karthikeyan S, Kathirvalavakumar T. Genetic Algorithm Based Over-Sampling with DNN in Classifying the Imbalanced Data Distribution Problem. *Indian Journal Of Science And Technology*. 2023;16(8):547–556. Available from: https://doi.org/10.17485/IJST/v16i8.863.
5) Liang D, Zhang JW, Tang YP, Huang SJ. MUS-CDB: Mixed Uncertainty Sampling With Class Distribution Balancing for Active Annotation in Aerial Object Detection. *IEEE Transactions on Geoscience and Remote Sensing*. 2023;61:1–13. Available from: https://doi.org/10.1109/TGRS.2023.3285443.

6) Wongvorachan T, He S, Bulut O. A Comparison of Undersampling, Oversampling, and SMOTE Methods for Dealing with Imbalanced Classification in Educational Data Mining. *Information*. 2023;14(1):54. Available from: https://doi.org/10.3390/info14010054.

7) Wang S, Dai Y, Shen J, Xuan J. Research on expansion and classification of imbalanced data based on SMOTE algorithm. *Scientific Reports*. 2021;11(1). Available from: https://doi.org/10.1038/s41598-021-03430-5.

8) Stefanowski. Dealing with Data Difficulty Factors While Learning from Imbalanced Data. *Studies in Computational Intelligence*. 2016;605:333–363. Available from: https://doi.org/10.1007/978-3-319-18781-5_17.

9) Yue L, Cai W, Cao D, Liu Y, Li Y, Wu J. Mitigate the Inter-channel Interference in Coherent Sampling-Based Nyquist OTDM Demultiplexer Using KNN Classifier. In: 2022 Asia Communications and Photonics Conference (ACP). IEEE. 2022;p. 935–938. Available from: https://doi.org/10.1109/ACP55869.2022.10088616.

10) Susan S, Kumar A. The balancing trick: Optimized sampling of imbalanced <scp>datasets—A</scp> brief survey of the recent State of the Art. *Engineering Reports*. 2021;3(4). Available from: https://dx.doi.org/10.1002/eng2.12298.

11) Mahmudah KR, Indriani F, Takemori-Sakai Y, Iwata Y, Wada T, Satou K. Classification of Imbalanced Data Represented as Binary Features. *Applied Sciences*. 2021;11(17):7825. Available from: https://doi.org/10.3390/app11177825.

12) Kumar P, Bhatnagar R, Gaur K, Bhatnagar A. Classification of Imbalanced Data:Review of Methods and Applications. *IOP Conference Series: Materials Science and Engineering*. 2021;1099(1):012077. Available from: https://dx.doi.org/10.1088/1757-899x/1099/1/012077.

13) Eldho KJ. Impact of Unbalanced Classification on the Performance of Software Defect Prediction Models. *Indian Journal of Science and Technology*. 2022;15(6):237–242. Available from: https://doi.org/10.17485/IJST/v15i6.2193.

14) Makki S, Assaghir Z, Taher Y, Haque R, Hacid MSS, Zeineddine H. An Experimental Study With Imbalanced Classification Approaches for Credit Card Fraud Detection. *IEEE Access*. 2019;7:93010–93022. Available from: https://dx.doi.org/10.1109/access.2019.2927266.