International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 12, Number 5 (2019), pp. 697-704

Optimization of kNN Classifier Using Hybrid Preprocessing Model for Handling Imbalanced Data

Preeti Nair[1]
Dept. of Computer Science and Engineering
Faculty of Engineering and Technology,
Manav Rachna International Institute of Research and Studies, Faridabad, 121004, India.

ORCID (0000-0002-5479-8703)

Indu Kashyap[2]
Professor, Dept. of Computer Sc. and Engineering
Faculty of Engineering and Technology,
Manav Rachna International Institute of Research and Studies, Faridabad, 121004, India.

ORCID (0000-0003-1884-0828)

Abstract

The conventional k Nearest Neighbor (kNN) classifier has many challenges when dealing with problems caused by imbalanced data sets. The classifiers are usually designed to improve accuracy by reducing the errors and therefore, they do not rely on class distribution or proportion or balance of classes. So for handling imbalanced data issues data preprocessing techniques such as sampling methods are widely used. One of them is Synthetic Minority Over-sampling Technique SMOTE. SMOTE with kNN was found to be deficient when it came to handling imbalanced data specifically in terms of accuracy. In this paper, we have proposed a preprocessing model for imbalanced datasets before it can be sent to the classifier. The proposed model is a hybridization of three preprocessing techniques; they are SMOTE, Resample technique and Attribute Selection technique combined to form SMOTE Resample Attribute Selection (SRAS). To see the performance of the proposed preprocessing hybrid model SRAS working with kNN classifier, we have taken several imbalanced datasets. These datasets have been tested and compared on standard kNN, SMOTE with kNN and SRAS with kNN. The proposed preprocessing model SRAS gave superior and enhanced outcomes for kNN classifier in terms of accuracy, recall, F Measure and AUC (area under the curve).

Keywords: Attribute selection, data preprocessing, Imbalanced datasets, kNN, Resample method, SMOTE.

I. INTRODUCTION

Classification is a widely used data mining technique. It is a process to assign an instance in a dataset to a target class. The objective of a classification is to accurately predict the class label for each unknown query instance. There are many classification techniques, which are widely used, some of them are Naive bayes, decision tree, kNN and many more. Here, our main focus is on kNN classification technique. kNN classifiers are based on learning by finding similarities between the instances. The similarities are measured by a distance formula. The least distanced instances are said to be similar or closer to each other. There are many distance measures such as Euclidean, Manhattan etc. that are used in calculating the similarity between two instances. When an unknown query instance i.e. whose class label is not defined, given to kNN classifier, it calculates distances from the unknown query instances to all the other data points. The value for k is provided and the classifier selects k similar or closest instances to the unknown instance. Then the majority classes in the k similar instances are observed and whichever is the majority class label is therefore assigned to the unknown query instance. For e.g. if majority class is "male class" then the query instance is assigned as "male". kNN classification algorithm is simple and easy to implement. kNN can analyse datasets expeditiously and which are usually handled by complex algorithms. [1][2].

In this research work, we are endeavouring to enhance the performance of kNN classification technique when working with imbalanced datasets.

Imbalanced datasets are those datasets in which the multitude of one class labels will be significantly more than another class labels. Like for e.g. if you have class labels in a datasets say A, B, and C, the ratio of the class label distribution will be uneven say 600:10:30, we can see the number of A is far larger than the other two labels. So, this is called imbalanced data. When an imbalanced data is classified, the results are overfit to A (as per our e.g. because A is in large number). So, accuracy is not just the performance criteria in the case of imbalanced datasets. Hence, we have to look for other measures as well, such as precision, recall, f-measure, AUC and so forth. There are many methods to resolve imbalanced datasets problem, some basic methods are sampling techniques like random under sampling, random over sampling, cluster based oversampling, SMOTE etc. these are the data level techniques to solve imbalanced datasets. There are some algorithmic ensemble technique as well like bagging and boosting to handle the imbalanced datasets.

In order to improve the performance of kNN classifier when working with imbalanced dataset, we have proposed a novel method which is a combination of two data level techniques called resample and SMOTE followed by attribute selection technique to form a hybrid model. The dataset produced by this model is then classified using kNN.

There are various previous studies related to Imbalanced data preprocessing. However some are done on one kind of data. For eg. Wosiak et al. have taken medical imbalanced dataset for this study. Here they have used various methods of balancing the datasets through data preprocessing step and applying it with different classification technique. The study shows that for some datasets there are certain combinations of preprocessing methods which when applied with certain classification technique would surpass other approaches. The resample preprocessing methods gave better outcomes such as accuracy and sensitivity. The study also shows that the hybrid approach of applying multiple preprocessing methods would surpass the single preprocessing methods. [3].

Additionally in another paper Abolkarlou et al. have proposed a hybrid method for preprocessing to handle the imbalanced data. In this method after using smote as the preprocessing technique, the second step is to determine the optimal layer after the data is separated into some layers using hierarchical clustering algorithm. Then Genetic algorithm is executed on three functions such as G-mean, diversity and G-mean * diversity. This hybrid method had been applied on some datasets and the results outperform the SMOTE bagging and boosting methods [4].

In a slightly different but similar approach Kang et al. have proposed a paper in which under sampling method is used as data preprocessing for imbalanced data. The minority class is taken in the training data and the minority samples sometimes contain noisy data which reduces the performance of the classifiers so, the authors have proposed a scheme in which noise filter is executed on minority samples before under sampling the majority class. Four under sampling methods are taken for applying this new scheme and the result specifies that the proposed scheme can improve the actual under sampling methods more significantly [5]. The impact of such an improvement in result is visible in another research paper. Wang et al. have proposed a novel preprocessing step for imbalance datasets by combining a data cleaning technique called tomek link technique with SMOTE. They have applied this technique with eight classifiers and three imbalance datasets to test the performance. The experimental results shows that the proposed system tomek link combined with SMOTE is significantly better than simply SMOTE combined with the classifiers [6].

Borowska et al. has created and tested two algorithms to improve the standard SMOTE algorithm. First, one is ASIS (Amplify SAFE Improved SMOTE). The second algorithm is named ADIS (Amplify DANGER Improved SMOTE). The experimental results showed that the proposed two algorithms produced much better results than the standard SMOTE [7]. Lopez et al. have discussed in this paper that the imbalanced ratio is not just the matter for poor performance of the classifier instead they say that skewed data is the main reason for it. Due to this finding, they have introduced some methods such as preprocessing of instances, cost effective learning and ensemble methods for imbalanced data. This review paper Fig 1. Shows the overview of the proposed model.

highlights that there is a need to study the skewed characteristics of the data such as the sparse sample size, overlapping class, noisy data, borderline management and the dataset shift, so that upcoming research on classification with imbalanced data can be enhanced [8].

The purpose of this research work is to find a novel hybrid preprocessing model which can improve kNN classifier when dealing with different imbalanced dataset. This research paper focuses on kNN because it is one of the most widely used classification algorithm and is relatively simple to use. Unlike the previous research work mentioned here which focused either on only one kind of dataset or discussed and worked with many different classifiers in general. The kNN is used in statistical estimation and pattern recognition and is particularly effective when the training data is large. This makes it more and more important in the future as we see the quantities of data generated in many such fields increasing geometrically. The efficiency of kNN being increased would therefore be a very desirable thing.

II. MATERIAL AND METHOD

The proposed model is prepared with three techniques, first one is SMOTE, and it's a well-known technique to confront the imbalanced dataset problem. It is an oversampling method which creates a synthetic sample for the minority class. SMOTE operates in feature space which results in generating synthetic samples in a less application specific mode. The line segment joining the k nearest neighbor along each minority samples is oversampled. The k nearest neighbour is randomly nominated depending upon the oversampling required [9].

The second model used to build the proposed model is the resample technique. In the resample technique a random subsample of dataset is produced using replacement or without replacement. The actual dataset must fit totally in memory. The number of instances may be specified in the produced dataset. The nominal class attributes is a must in the dataset otherwise unsupervised version is used. A filter can be made in order to bias distribution of the class or to maintain the subsample distribution of class [10].

The third model is Attribute selection. Attribute selection is also called feature selection is a method that automatically selects those attributes from the datasets which are most relevant for constructing predictive models. Attribute selection methods include and exclude attributes present in the data without changing them. Attribute selection methods helps in creating an accurate analytical model by selecting features that will give better accuracy using less data. The unwanted, inappropriate and superfluous attributes from data that do not contribute to the accuracy of an analytical model or may in fact reduce the accuracy of the model are removed by this method therefore reducing the complexity of the model and it is effortless to comprehend. Hence, it helps in creating a more cost effective model [11].

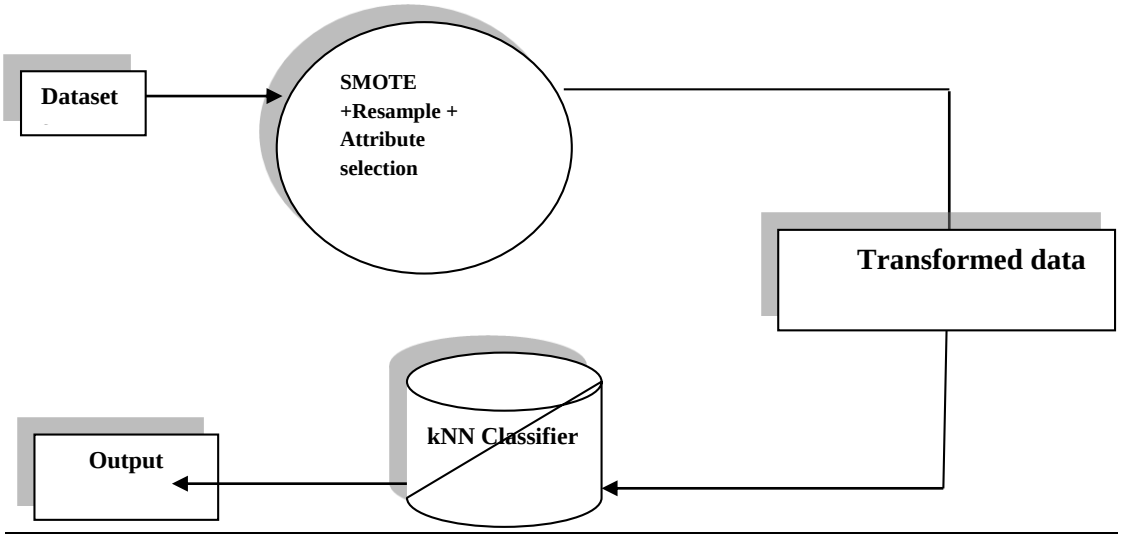

Fig 1. Proposed Hybrid preprocessing model Flow Diagram.

The objective of this paper is to enhance the performance of kNN classifier when handling or managing imbalanced data. So we have proposed a preprocessing step before the classification process. In this preprocessing step, we have created a hybrid model called SRAS which can handle imbalanced datasets. This proposed model has two processes within it, first it transforms the dataset into a balanced dataset using the combination of SMOTE + Resample technique. Then it selects important attributes responsible for accurate prediction by the means of attribute selection technique. To examine the performance of the proposed model several benchmark datasets from KEEL and UCI data repository has been considered. Table 1. Describes the algorithm for the proposed model. Table 2. Shows the list of datasets taken for this paper.

Table 1. Algorithm for the Proposed Hybrid Preprocessing Model SRAS for balancing Imbalanced Datasets before kNN Classification.

Algorithm 1 :Hybrid Preprocessing Model SRAS
Input: Imbalanced Dataset **Output: Balanced dataset with selected attributes, kNN classification result** **Step 1:** **1: Load the dataset (D_x) for preprocessing** **2: Build the hybrid preprocessing model: SMOTE and Resample and Attribute Selection.** **3: Transformed dataset $(D_x)'$ with selected attributes is obtained.** **Step 2:** **Apply $(D_x)'$ on kNN classifier** **Output: classification result.**

II.I Datasets Used

Some of the well-known imbalanced datasets are used in this paper. The imbalanced datasets taken here are both dichotomous as well as multi label datasets. [12][13].

Table 2. Description of the imbalanced datasets used in this study.

Datasets	Instances	Attributes	Labels	Imbalanced Ratio
Glass	214	10	7	1.82
Dermatology	358	34	4	16.9
Page Block	5472	10	5	8.79
Vehicle	846	18	4	3.25
Ecoli	336	8	8	3.36
Heberman	306	3	2	2.78
Thoracic Surgery	470	17	2	5.71

II.II Classification Performance Evaluation Measures and Formulae

In order to evaluate the classification performance, we calculate the count of True Positive, False Negative, False Positive and True Negative.

True positive implies to the positive instances that were correctly labelled as positive by the classifier. True negative implies to the negative instances that were correctly labelled as negative by the classifier. False positive denotes the negative

instances that were incorrectly labelled as positive. False negative denotes positive instances that were mislabelled as negative [1] [2].

• Accuracy is the overall accuracy of the classifier and it is formulated as

$$Accuracy: \frac{TN + TP}{N + P} \quad (1)$$

• Sensitivity (Recall) Recall measures the rate of positives correctly predicted as positive by the classifier. It is formulated as

$$Recall: \frac{TP}{P} \quad (2)$$

• Precision: Precision measures the correctness rate of the class predictions done as positive by the classifier.

$$Precision: \frac{TP}{TP + FP} \quad (3)$$

• F-Measure: the harmonic mean of precision and recall is the F-Measure. It can also be regarded as the weighted average of recall and precision.

$$Fmeasure: 2 \times \frac{Recall \times Precision}{Recall + Precision} \quad (4)$$

• Imbalanced Ratio (IR): The imbalance ratio is computed by taking the division of a number of instances in the majority class to the number of instances in the rare class.

$$IR: \frac{Numberofnegativeclassinstances}{Numberofpositiveclassinstances} \quad (5)$$

• Area under Curve (AUC): The fraction of whole area that comes under the ROC graph is called AUC curve. For classification performance evaluation, AUC provides a single value and ROC is computed for overall measure of quality. AUC can also be used for evaluating the performance of Imbalanced datasets. The formula for AUC is

$$AUC: \frac{TP_{rate} + TN_{rate}}{2} \quad (6)$$

III. RESULTS AND DISCUSSIONS

The experiment has been performed individually on each datasets. First the datasets where executed using the existing kNN without preprocessing, Secondly the datasets was executed on SMOTE with kNN, thirdly it was executed on the proposed pre-processing model (SRAS) with kNN. The results of all the three models are compared in terms of performance evaluation criteria such as accuracy, precision, recall, FMeasure and AUC. The experiments are performed on WEKA tool [10].

Table 3. Shows the Accuracy values of the new proposed system versus the two well-known existing models kNN, SMOTE+kNN, and the bold values represent the highest values obtained during the experiments done individually on each model with respect to the datasets. Fig 2 is the corresponding graph to the Accuracy values showed in the Table 3. The light blue in the graph shows the highest value which is obtained by the proposed system SRAS+kNN. Table 4. and Fig 3, Fig 4 and Fig 5 describes about the classification performance values such as Precision, Recall and FMeasure obtained when the datasets were executed on the three models, the grey in the graphs shows the higher values obtained. The bold values in all the tables shows the highest values obtained. Higher values signify that the classification results are more accurate than the other two models. Table 5. Describes the AUC values of the three models. The proposed model outperforms the other two models significantly.

Table 3. Accuracy values of conventional kNN classifier without preprocessing, SMOTE and kNN and the proposed preprocessing SRAS with kNN.

Datasets	kNN	SMOTE+kNN	SRAS+KNN
glass	.7056	.7040	**.9148**
dermatology	.9454	.9456	**.9896**
page Block	.9602	.9616	**.9815**
vehicle	.6986	.7608	**.8823**
Ecoli	.8036	.8047	**.9408**
Heberman	.683	.7054	**.8630**
Thoracic Surgery	.7723	.7019	**.8833**

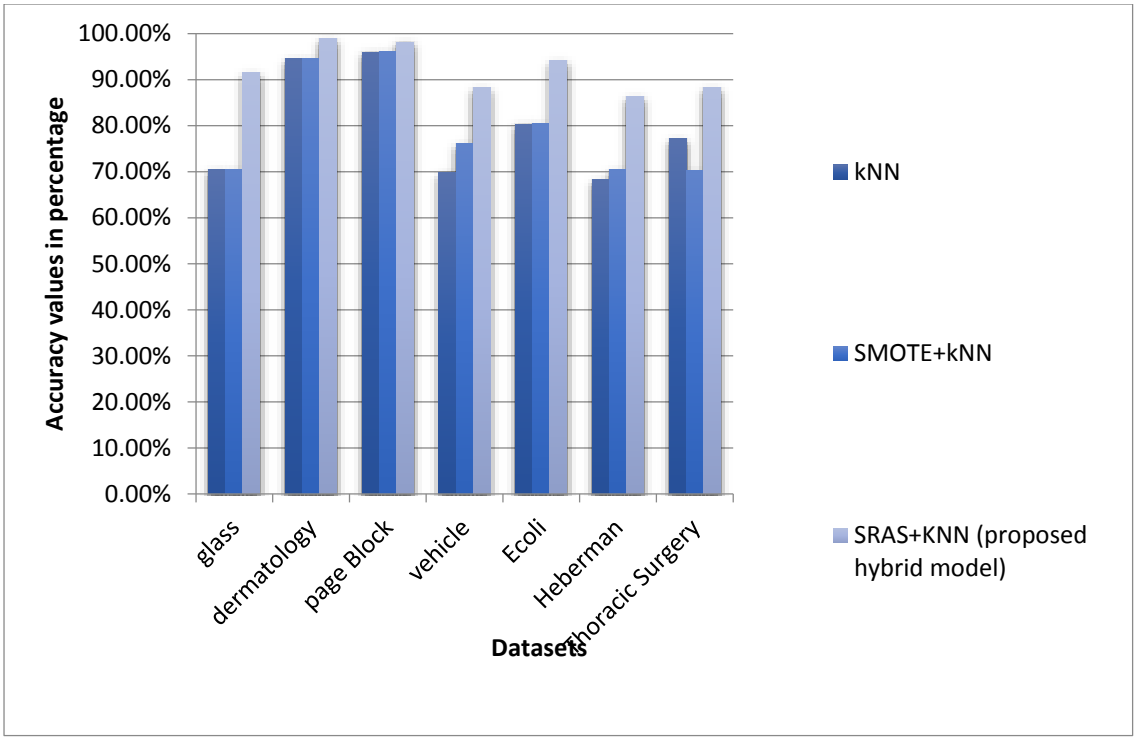

Fig 2. This graph represents the Accuracy values of kNN, SMOTE+kNN and SRAS+kNN (proposed model)

Table 4. Precision, Recall and Fmeasure Values of conventional kNN classifier without preprocessing, (SMOTE+kNN) and the Proposed preprocessing (SRAS+ kNN) Model

	Precision			***Recall***			***FMeasure***		
Datasets	**kNN**	**SMOTE+kNN**	**SRAS+kNN**	**kNN**	**SMOTE+kNN**	**SRAS+kNN**	**kNN**	**SMOTE+kNN**	**SRAS+kNN**
Glass	.709	.704	**.917**	.706	.704	**.915**	.704	.702	**.915**
Dermatology	.947	.947	**.990**	.945	.946	**.990**	.945	.945	**.990**
Page Block	.959	.960	**.981**	.960	.962	**.981**	.959	.961	**.981**
Vehicle	.691	.748	**.882**	.699	.761	**.882**	.694	.754	**.882**
Ecoli	.799	.799	**.941**	.804	.805	**.941**	.801	.801	**.940**
Heberman	.653	.704	**.863**	.683	.705	**.863**	.664	.705	**.863**
Thoracic Surgery	.754	.707	**.882**	.772	.702	**.883**	.763	.704	**.882**

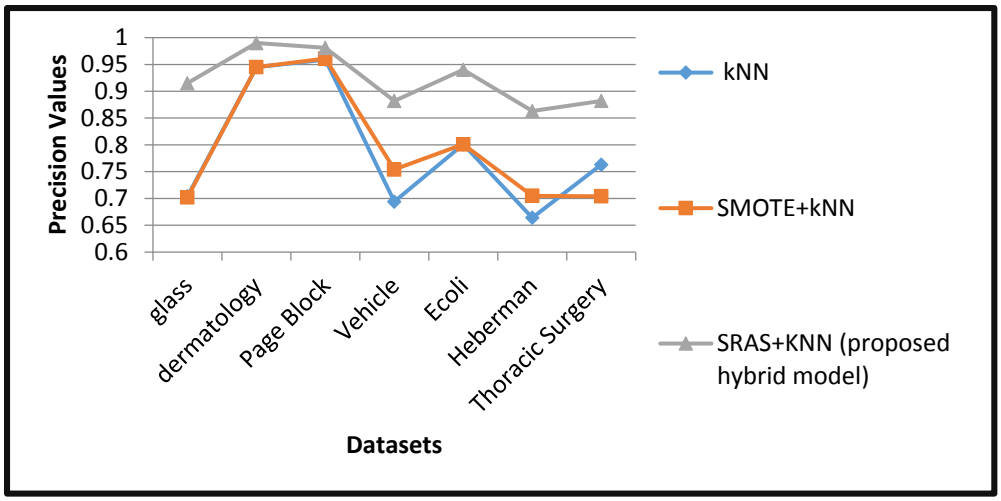

Fig 3. Precision values of kNN worked without preprocessing, kNN and SMOTE and the proposed hybrid model SRAS with kNN

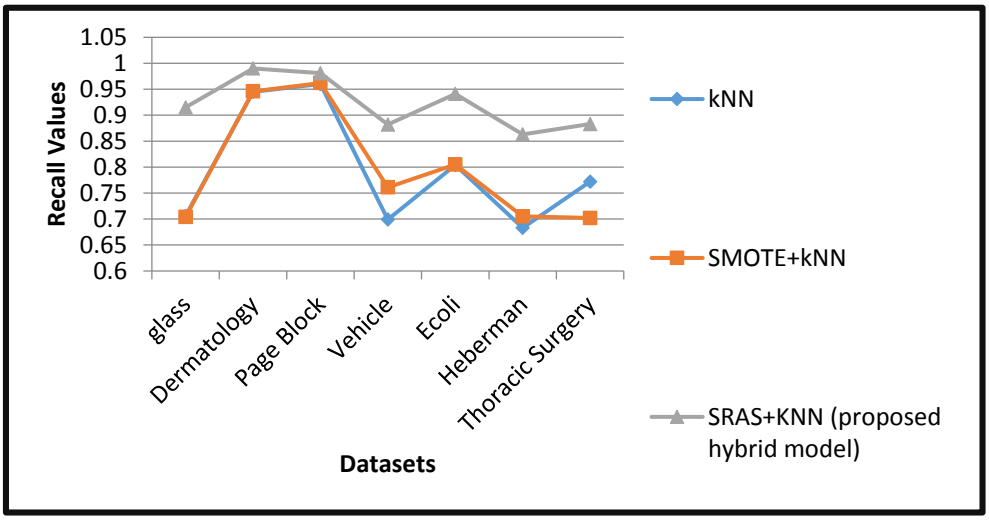

Fig 4. Recall values of kNN worked without preprocessing, kNN and SMOTE and the proposed hybrid model SRAS with kNN

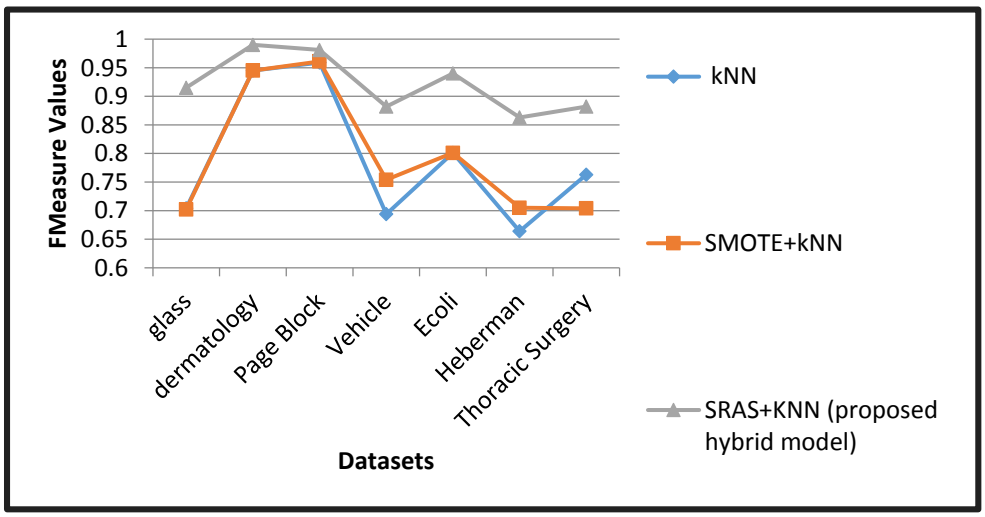

Fig 5. FMeasure values of kNN worked without preprocessing, kNN and SMOTE and the proposed hybrid model SRAS with kNN

Table 5. AUC-ROC values of conventional kNN classifier without preprocessing, kNN after preprocessing SMOTE and kNN after the proposed preprocessing SRAS

Datasets	kNN	SMOTE+kNN	SRAS+kNN (proposed hybrid model)
Glass	.792	.801	**.949**
Dermatology	.969	.972	**.999**
Page Block	.880	.880	**.949**
Vehicle	.802	.848	**.921**
Ecoli	.878	.878	**.961**
Heberman	.577	.732	**.865**
Thoracic Surgery	.513	.620	**.854**

III. CONCLUSION

An attempt has made to boost the performance of kNN classification algorithm when handling imbalanced datasets. When classifying imbalanced datasets there is a problem of high false negative rates, so to decrease false negative rates and to enhance the accuracy, the dataset balanced has to be balanced. Even with a well-trained classifier like kNN, we cannot reduce the false negative rates. In this paper, a novel hybrid preprocessing model has been proposed which transforms the imbalanced datasets to a balanced dataset by applying the hybrid combination of two sampling techniques SMOTE and resample along with Attribute selection (SRAS) before it is passed on to the kNN classifier. For this, several imbalanced datasets have been considered for testing the performance of the proposed model (SRAS+kNN). The imbalanced datasets were executed individually on standard kNN, SMOTE with kNN and on our proposed model SRAS with kNN. Results show that much enhanced outcomes were observed for the proposed system in terms of classification performance measurements such as precision, recall, AUC, FMeasure and accuracy.

As medical research increases in scope and breadth, the proclivity of data to be biased and greatly magnified increases almost geometrically. This situation makes the ability of tools to handle such data literally lifesaving in the long run. The research done within the scope of this paper is to develop one such tool and thus has great significance in the future. Another major use of this proposed model is that it can be embedded with other classification models thereby enhancing their proficiency as well.

IV. ACKNOWLEDGEMENT

This work would not have been possible without the help and guidance of my supervisor, the head of the department of the Faculty of Engineering and Technology (FET), and other Faculty of the FET of Manav Rachna International Institute of Research and Studies (MRIIRS).

REFERENCES

[1]. J. Han, M. Kamber, and J. Pei, Data Mining Concepts and Techniques, Waltham, MA, USA: Third edition, Morgan Kaufmann,(2012).

[2]. P. Nair, N. Khatri, and I. Kashyap, A novel technique: ensemble hybrid 1NN model using stacking approach, International Journal of Information Technology, Springer (2018), https://doi.org/10.1007/s41870-018-0109-0.

[3]. A. Wosiak, and S. Karbowia, Preprocessing compensation techniques for improved classification of imbalanced medical datasets, in Federated Conference on Computer Science and Information Systems (FedCSIS) 2017 DOI: 10.15439/2017F82

[4]. N. Abolkarlou, and K. Ebrahimpour, Ensemble imbalance classification: Using data preprocessing, clustering algorithm and genetic algorithm, in 2014 4thInternational Conference on Computer and Knowledge Engineering (ICCKE) DOI: 10.1109/ICCKE.2014.6993364

[5]. Q. Kang, X. Chen, S. Li, and M. Zhou, A Noise-Filtered Under-Sampling Scheme for Imbalanced Classification, IEEE TRANSACTIONS ON CYBERNETICS,(2017). Year: 2017 , Vol: 47 , Issue: 12,Pages: 4263 – 4274

[6]. L. Wang, M. Zeng, B. Zou, W. Faran, and X. Liu, Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data, in ICOACS, 2016. 10.1109/ICOACS.2016.7563084

[7]. K. Borowska K, and M. Topczewska, Data preprocessing in the classification of the imbalanced data, Advances in Computer Science Research, 2014. vol. 11, pp. 31-46.

[8]. V. Lopez, A. Fernandez, S. Garcia, V. Palade, F. Herrera, An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics, Information Sciences, ELSEVIER 2013. http://dx.doi.org/10.1016/j.ins. 2013.07.007.

[9]. N. Chawla, B. Kevin, L. Hall and P. Kegelmeyer, SMOTE: Synthetic Minority Over-sampling Technique, Journal of Artificial Intelligence Research 16, (2002) 16(1):321-357.

[10]. R. R. Bouckaert, E Frank, M. Hall, R. Kirkby, P. Reutemann, A. Seewald and D. Scuse. WEKA Manual for Version 3-8-0, University of Waikato, Hamilton, New Zealand April 14, 2016.

[11]. I. Guyon, and E. Andre, An Introduction to Variable and Feature Selection, Journal of Machine Learning Research 3 (2003) 1157-1182.

[12]. UCI machine learning repository. [online]. Available: http://archive.ics.uci.edu/ml 2016.

[13]. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. García , L. Sánchez and F. Herrera, KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework, Journal of Multiple-Valued Logic and Soft Computing 17:2-3, (2011) 255-287.