iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 10 **Issue:** X **Month of publication:** October 2022

DOI: https://doi.org/10.22214/ijraset.2022.47218

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 10 Issue X Oct 2022- Available at www.ijraset.com

A Survey on Cardiac Disease Classifying from Imbalanced Healthcare Facts Using Ensemble Classification Techniques

R. Saranya[1], Dr. D. Kalaivani[2]
[1]*MCA, M.Phil., Ph.D [Part Time] Research Scholar, Department of Computer Science, Dr.SNS Rajalakshmi College of Arts and Science (Autonomous), Coimbatore-49, Tamil Nadu, India.*
[2]*Ph.D, Associate Professor & Head, Department of Computer Technology, Dr.SNS Rajalakshmi College of Arts and Science (Autonomous), Coimbatore-49, Tamil Nadu, India.*

Abstract: ***Heart disease is the leading problem of death globally, and early detection is pivotal in preventing the progression of the disease. In this paper, Improved EnsembleSVM method is proposed for the prediction of heart disease risk. The technique involves randomly partitioning the dataset into smaller subsets using a mean based splitting approach. Oversampling of imbalanced data introduces the extension of Synthetic Minority Over-sampling Technique through a recent ideology, and recurrent ensemble-based noise filter called*** **duplicative-Partitioning Filter,** ***which can overwhelm the hindrance fashioned by noisy and frontier models in overbalanced dataset. Ensemble classifier with oversampling technique plays an accurate result for predict heart disease in efficient way.***
Keywords: Classification, Machine Learning, Preprocessing heart disease data using Improved EnsembleSVM

I. INTRODUCTION

Cardiac disease where coronary arteries that supply oxygen rich blood to the heart muscles are blocked. This affects the normal functioning of the heart.

Symptoms for this include chest pain, discomfort in shoulder and back, fatigue, shortness of breath, nausea, heart burn and finally heart attacks or also known as angina. 9, expressly sampling, progressive learning, estimate tactful learning and kernel based technique. Sampler based tactics support the remedy at data level by interrelating the number of fragments among edification. Undersampling and oversampling are twin key species of selection in which snippets are either diminished from major part grade or selections are top-up in the faction class. The pair approaches have their specific privileges as well as obstacles. Exertive intellect techniques core essentially on securing brands to the unlabeled data.

Sampling methods based on data are usually simple and intuitive. Undersampling method usually causes information loss while oversampling method tends to balance the original dataset. Thus, the latter one is often adopted in data classification. The proposed work introduces novel algorithm called H-SMOUTE, which is done a critical modification to Synthetic Minority Oversampling Technique (SMOTE) for highly imbalanced datasets, where the generation of new synthetic samples are directed closer to the minority than the majority. In this way, the line of distinction between the two classes will be clearly defined and all samples in data will be located within their class boundaries to ensure accurate prediction of the classifiers developed.

An ensemble is fabricated in two approaches, i.e., producing the support learners, and then combinative them. Support learners are habitually generated from training data by a support learning procedures which can be decision tree, neural network or other kinds of deep learning subroutines. Ensemble techniques have previously gained tremendous eminent in multiple real-world tasks, such as medicative diagnosis and remote grasping.

II. RELATED WORKS

Sayan Surya Shaw et al[16].,Most of the disease datasets, prepared by some means, are imbalanced in nature, which implies that number of instances belonging to one class (i.e. the minority class) is exceptionally less compared to the number of instances in the other class (i.e. the majority class). Hence, if we directly feed such data to a classification model, it would mislead the model performance.

III. CLASSIFICATION OF IDS

Imbalanced Data Sets (IDS), also mentioned to as class unbalance learning, agree to preserve where there are a huge amount of paradigms of some classes than others. Grouping on IDS habitually premises issues because standard deep learning set of rules gravitate to be overwhelmed by the immense groups and neglect the small ones. Most classifiers engage on data exhausted against the unique propagation as the training data, and imagine that maximizing factuality is the principle goal. Imbalanced Data Sets (IDS) problem, also known as class discrepancy problem, effectively corresponds to the obstacle distinguished by primitive learning set of rules on empires for which some grades are illuminated by a massive notation of instances while others are portrayed by only a few[9]. We normally meet two-class hurdles, which mean one class has much more instances than the other. Erratically, we also have multi-class cases, in which there are not lavish instances for more than one class. It may cause more trouble when decisive classification boundary. Classification models reveal low accuracy when dealing with imbalanced datasets. Therefore, a number of models will be evaluated with the objective to find those that better address the classification problem of electronic health records of imbalanced datasets[1]. A special focus will be given to the investigation of some ways of swapping with electronic health records based unbalanced datasets lying on a Random Forest.

The crucial origins of the proposed role are as follows:

1) The scrutiny of current obstacles of medical overbalanced learning;
2) The investigation of dissimilar grouping models and estimation metrics for imbalanced datasets;
3) The investigation of grouping models based on Random Forest;
4) The comparison of different grouping models for medical overbalanced datasets.

As a chunk of my research, ensemble is designated as a probably effective way to decode class imbalance obstacles. An ensemble of classifiers is a dump of classifiers whose specific decisions are collated in some way to categorize new models[11]. Each algorithm takes an initiator and a training set as input and runs the learner numerous times by fluctuating the propagation of training set instances. The promoted classifies are then collated to create a final classifier that is inherited to classify the experiment set.

IV. MACHINE LEARNING

One of the most well-known machine learning algorithms tasks is the classification of data. Machine learning tends to be an essential function in this case for extracting knowledge from business activity datasets and transferring it to larger databases. The majority of the machine learning methods rely on a huge number of features that explain the algorithm's behavior, resulting in the model's complexity, indirectly or directly [10]. Many algorithms such as hybrid methods are used in conjunction with logistic regression, naive bayes, k-nearest neighbor, and neural networks to integrate the heart disease diagnostic algorithms mentioned earlier. Thus, in this case, the system was trained and implemented over the python platform with the help of the uci (unique client identifier) machine learning deported benchmark dataset.

Coronary artery disease, arrhythmias (heart rhythm problems), heart abnormalities (such as congenital heart defects), and a variety of other disorders are included in the category of heart diseases. Cardiomyopathy and heart infections are among the conditions that fall under this category. The most common measure of heart risk is chest pain, which is a symptom of cardiovascular disease. After that, it has symptoms of Nausea, Indigestion, Heartburn, or Stomach Pain. The paper will exhibit how a program can be created in Python to analyze whether or not an individual is suffering from cardiovascular disease or not [11]. In this paper, the system uses a dataset comprising fourteen characteristics of the test outcomes, carried on around 100 persons. However, the patient suffering from heart disease symptoms will be diagnosed using binary digits, 1 and 0, where 1 will indicate the true value (The patient has heart disease, in other words.) And 0 will indicate the false value (that is, the patient does not have any kind of heart disease). Additionally, co-relation and trends of the obtained features will also be recognized with the help of several features, such as gender, age, cp (chest pain type), chol (cholesterol level), FBS (fasting blood sugar level), exang (exercise-induced angina), thalach (maximum achieved heart rate), old peak (ST depression persuaded by exercise respective to rest), thal (maximum achieved heart rate), ca (number of major vessels).

V. PREPROCESSING HEART DISEASE DIAGNOSIS USING ISMOTE

Generally, many of the Machine learning algorithms applied to classification problems assume that the classes are well balanced. Data sampling is the most common method used to solve the class imbalance problem [5]. The data sampling method involves creating a balanced dataset by adjusting the number of samples of the majority class, which occupies most of an imbalanced dataset, and the minority class, which occupies a small part. The sampling method can be classified as an Undersampling or oversampling method depending on for which of the two classes the number of samples is adjusted [12]. The Synthetic Minority Oversampling Technique (SMOTE) is an oversampling process achieved using additional synthetic data.

According to, the original data obtained using SMOTE is used to synthesize new minority data that are different from the original ones, thereby alleviating the impact of overfitting on the minority class.

The SMOTE is based on the idea of the nearest neighbor algorithm (kNN) and assumes that a synthetic data sample can be interpolated between an original and one of the closest neighbors. The SMOTE algorithm calculates the neighbor environment of each data sample from the minority class randomly selects one of its neighbors and makes synthetic data through the interpolation of data between each sample and the nearest neighbor selected. When the number of synthetic data samples to be made is smaller than the size of the original dataset, the algorithm is randomly selected and an original data sample is used to create synthetic data samples. Conversely, when the number of synthetic data samples to be made is greater than the size of the original dataset, the algorithm iteratively creates synthetic data samples using predetermined.

The SMOTE algorithm is described in detail below:

1) Find the k-nearest neighbors for each sample.
2) Select samples randomly from a k-nearest neighbor.
3) Find the new samples = original samples + difference * gap (0, 1).
4) Add new samples to the minority. Finally, a new dataset is created.

The SMOTE method comes with some weaknesses related to its insensitive oversampling where the creation of minority samples fails to account for the distribution of sample from the majority class [17]. This may lead to the generation of unnecessary minority samples around the positive examples that can further exacerbate the problem produced for borderline and noisy in the learning process.

A. Improved-Synthetic Minority Oversampling Technique (I-SMOTE)

To perform better prediction, most of the classification algorithms strive to obtain pure samples to learn and make the borderline of each class as definitive as possible. The synthetic examples that are far away from the borderline are easier to classify than the ones close to the borderline, that pose a huge learning challenge for majority of the classifiers. On the basis of these facts, here present a new improved approach (I-SMOTE) for preprocessing of imbalanced training sets, which tries to clearly define the borderline and generate pure synthetic samples from SMOTE generalization [15].

Our proposed method has two stages and discussed as follows:

1) First stage, Here, first apply SMOTE algorithm to generate the synthetic instance based on following equation:

$$N = 2 * .r - z/ + z \quad \text{(1)} \qquad \text{… equ. (1)}$$

Where N is the initial synthetic instance number (newly generated), r, is the number of majority class samples, and z, is the number of minority class samples.

2) Second stage, To eliminate the synthetic samples with higher proximity to the majority class than the minority as well as the synthetic instances closer to the borderline generated by SMOTE. The A-SMOTE procedure step-by-step is outlined as follows:

a) *Step 1:* The synthetic instances that generated by SMOTE might be accepted or rejected on two conditions and it matches with the first stage:
b) *Step 2:* After that, with the accepted synthetic instances the following is carried out to eliminate the noisy.
c) *Step 3:* Similarly, to calculate the distance.

VI. EXPERIMENTAL STUDY

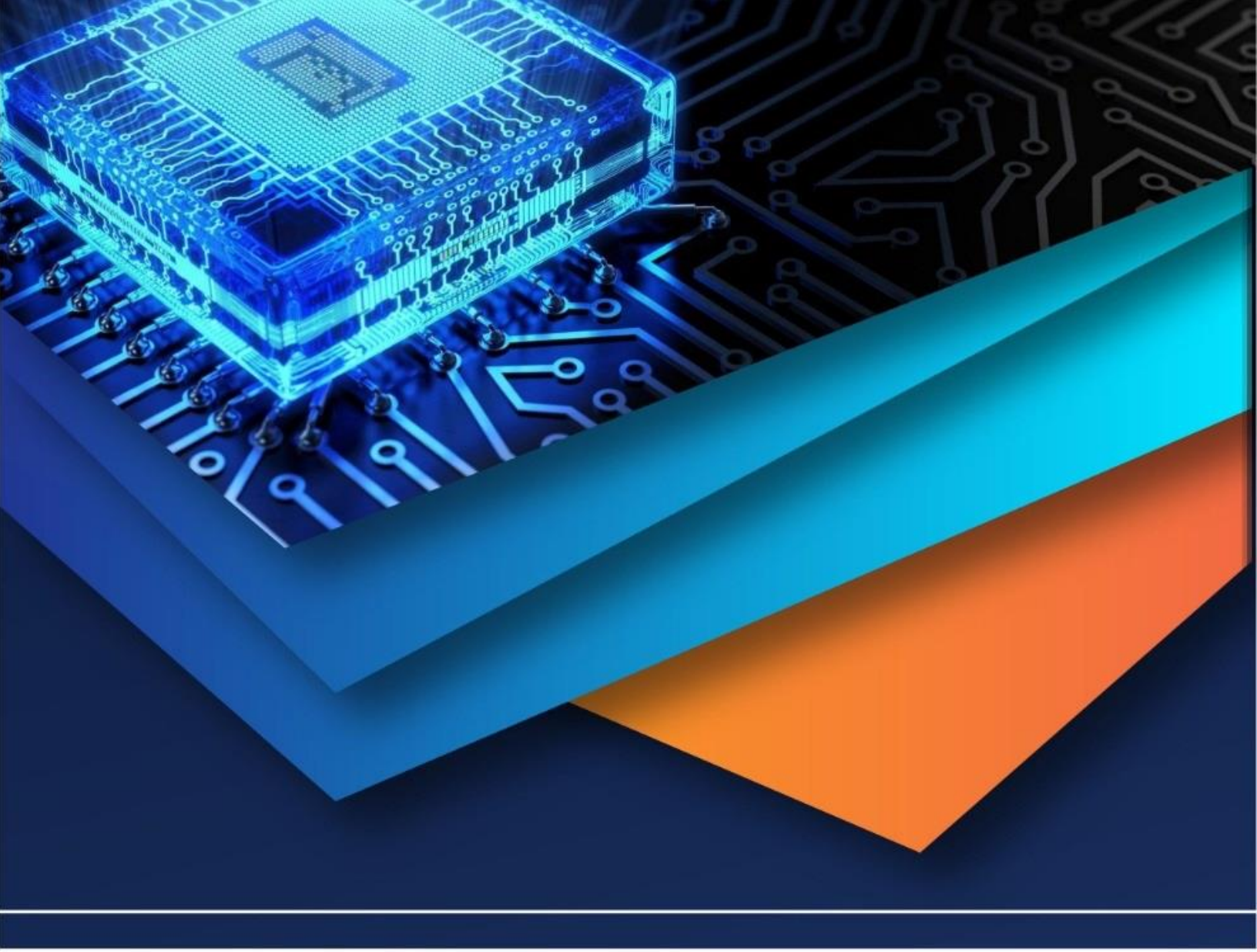
1) In this part, present the experimental design and the results based on the evaluation metrics employed, datasets, different imbalanced methods, and statistical tests. The experiments carried out using MATLAB (2016a). In this research, illustrate the datasets used for the experimental study and the statistical tests used alongside the empirical analysis. The proposed work has used 44 datasets from the KEEL data repository with highly imbalanced rates.
2) The evaluation criterion is a key factor both in the assessment of the preprocessing performance and guidance of the classifier modeling. In a two-class problem, the electronic health records the results of correctly and incorrectly recognized examples of each class.

VII. CONCLUSION

In this paper, Imbalanced data sets (IDS) problem plays a vital task in the healthcare system of the world, and the most vital feature is that the investigative results directly affect the long-suffering's treatment and safety. This research first balances the data size of each class by reducing the data in the majority class and adding virtual samples to the minority one. In this study, proposed a novel approach for highly imbalanced datasets, I-SMOTE, this is an improvement on SMOTE techniques. The proposed I-SMOTE can be a useful tool for researchers and practitioners since it results in the generation of high-quality data. Hence, this paper is mainly to analyze and predict the heart disease diagnosis using ISMOTE techniques.

REFERENCES

[1] Eshtay, M., Hm Faris., & N, Obeid. "Improving Extreme Learning Machine by Competitive Swarm Optimization and its application for medical diagnosis problems", Expert Systems with Applications", 104, 134-152, 2018.
[2] Shen, L., Chen, H., Yu, Z., Kang, W., Zhang, B., Li, H., Yang, B.,& Liu, D. "Evolving support vector machines using fruit fly optimization for medical data classification", Knowledge-Based Systems, 96, 61-75, 2016.
[3] Liu, Y., Yu, X., Huang, J X.,"Combining integrated sampling with SVM Ensembles for learning from imbalanced datasets", Information Processing & Management, 47(4),617-631, 2011.
[4] Papouskova, M., Hajek, P. "Two-stage consumer credit risk modeling using heterogeneous ensemble learning", Decision Support Systems, 118, 33-45, 2019.
[5] Onan, A., S, Korukoğlu., & H, Bulut. "A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification" Information Processing & Management, 53(4), 814-833, 2017.
[6] Na Liu,, Xiaomei Li1, Ershi Qi1, Man Xu, Ling Li And Bo Gao,"A novel Ensemble Learning paradigm for Medical Diagnosis with Imbalanced Data", 10.1109/ACCESS.2020.3014362,2017.
[7] K. Ravi, and V. Ravi. "A novel automatic satire and irony detection using ensembled feature selection and data mining", Knowledge-Base Systems, 2017.
[8] J. Luengo, A. Fern´andez, S. Garc´ıa, and F. Herrera, "Addressing data complexity for imbalanced data sets: analysis of SMOTE basedoversampling and evolutionary undersampling," Soft Computing, vol. 15, no. 10, pp. 1909–1936, 2011.
[9] Y. Tang, Y.-Q. Zhang, and N. V. Chawla, "SVMs modeling for highly imbalanced classification," IEEE Transactions on Systems,Man, and Cybernetics B: Cybernetics, vol. 39, no.1, pp.281–288,2009.
[10] YanWei, Ni Ni, Dayou Liu, Huiling Chen, MingjingWang, Qiang Li, Xiaojun Cui, and Haipeng Ye, "An Improved Grey Wolf Optimization Strategy Enhanced SVM and Its Application in Predicting the Second Major", Hindawi Mathematical Problems in Engineering Volume 2017.
[11] Xu Z , Shen D , Nie T , et al. A hybrid sampling algorithm combining M-SMOTE and ENN based on Random Forest for medical imbalanced data. Journal of Biomedical Informatics, 2020.
[12] Raghuwanshi, B.S. and S. Shukla, SMOTE based class-specific extreme learning machine for imbalanced learning. Knowledge-Based Systems, 2019.
[13] Douzas, G., F. Bacao and F. Last, Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Information Sciences, 465: p. 1-20, 2018.
[14] I. H. Witten, E. Frank, M. A. Hall, C. J. Pal, Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann, 2016.
[15] S. Oh, M. S. Lee, and B. T. Zhang, "Ensemble learning with active example selection for imbalanced biomedical data classification," IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 2, pp. 316–325, 2011.
[16] Sayan Surya, Shameem Ahmed, Samir Malakar, Ram Sarkar, An Ensemble Approach for Handling Class Imbalanced Disease Datasets, Proceedings of International Conference on Machine Intelligence and Data Science Applications pp: 345-355 May 2021.
[17] Verbiest, N., Ramentol, E., Cornelis, C., & Herrera, F.(2014).Preprocessing noisy imbalanced datasets using SMOTE enhanced with fuzzy rough prototype selection. Applied Soft Computing, 22, 511-517.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY