International Journal of Computer Techniques – Volume 11 Issue 2, 2024

RESEARCHARTICLE OPENACCESS

A Combinatorial AWS Cluster Cloud and Swarm Method for Unbalanced Classification.

ARUPULLA MAMATHA [1] , NARAYANADAS ARPITHA [2] , Prof. CH.G.V.N Prasad[3]
[1,2](Assistant professor, Dept of CSE ,Sree Dattha Institute of Engineering and Science).
[3] (Professor, Dept. of CSE, Sri Indu College Of Engineering & Technology (A)).

Abstract- This work presents an ensemble approach for multi-class imbalanced data classification called SSO-Adaboost-KNN . This paper's primary objective is to combine feature selection and boosting into an ensemble and increase the minority class's accuracy rate compared to the current technique for classifying imbalanced data. First, a feature selection technique with fitness-based weight adjustment is suggested in this model using Simplified Swarm Optimization (SSO). Secondly, Adaboost use the KNN technique to reduce the weights of the majority class that are in close proximity to the minority class, allowing the classifier to focus more on the minority class. The outcomes demonstrate that, following feature selection, the suggested approach performs better and increases boosting's accuracy and stability.

Key Index Terms— Machine learning, SSO, ensemble learning, and imbalanced data.

I.INTRODUCTION

The biggest concern in data mining is class inequality. These days, imbalanced data can be a hindrance to data mining; minority classes can have greater significance than majority classes in certain situations, such as credit card fraud, medical diagnosis, and other cases. Because of their infinite quantity and imbalance, data becomes challenging to classify. When one of the two classes has a larger sample size than the other, an imbalance problem arises. Most algorithms concentrate primarily on classifying the major sample, neglecting or incorrectly classifying the minority sample. The very significant but infrequently occurring minority samples are known as such. The imbalance data set can be classified using a variety of techniques . Learning from unbalanced data presents a problem since the relatively or

Without a doubt The learning algorithm cannot receive equal attention from underrepresented classes. In contrast to the majority class, this frequently results in the minority class having very specific classification rules or no rules at all, with no capacity for future prediction [2]. One of the main research questions in class imbalance learning is how to identify minority class data more effectively. Getting a classifier that will provide the minority class high accuracy without seriously endangering the accuracy of the majority class can be a general description of its learning objective [3]. In data mining, data sampling has drawn a lot of attention in relation to the issue of class imbalance. By adding samples to or removing samples from the data set, data sampling attempts to address the issue of uneven class distributions [5].

II.LITERATURE REVIEW

This strategy increases the minority class's classification accuracy, however it is not appropriate for classifying skewed data streams due to infinite data streams and constant idea drifting. The majority of imbalance learning strategies now in use are limited to two class problems. Class decomposition is mostly used to tackle multiclass imbalance problems. An ensemble learning algorithm called AdaBoost.NC [4] combines the boosting technique with the strength of negative correlation learning. The primary use case for this approach is multiclass imbalance datasets. Compared to other existing class imbalance learning approaches, the results imply that AdaBoost.NC paired with random oversampling can increase the prediction accuracy on the minority class without sacrificing overall performance. The algorithm for categorization was proposed by Wang et al.

for skewed data stream in [5], which shows that clustering sampling outperforms the traditional under sampling, since

Clustering helps to reserve more useful information. However, the method cannot detect concept drifting. Chris [3] proposed that both sampling and ensemble technique are effective for improving the classification accuracy of skewed data streams. SVMbased one-class Skewed data streams learning method was proposed in [7].

One of the most common data sampling techniques is Random Under-sampling**.** RUS simply removes examples from the majority class at random until a desired class distribution is achieved. RUSBoost is a new hybrid sampling and boosting algorithm for learning from skewed training data. RUS Boost provides a simpler and faster alternative to SMOTE Boost which is another algorithm that combines boosting and data sampling [8]. RUS decreases the time required to construct a model, which is benefit when creating an ensemble of models [9] that is use in boosting. RUS Boost presents a simpler, faster, and less complex than SMOTE Boost for learning from imbalanced

data. SMOTEBoost combines a popular oversampling technique (SMOTE) with AdaBoost, resulting in a hybrid technique that increases the performance of its components. Infinitely imbalanced logistic regression [10] a recently developed classification technique that is named infinitely imbalanced logistic regression (IILR) acknowledges the problem of class imbalance in its formulation and this technique modify the distribution of training data such that

cost of example calculated based on appearance of Example. Threshold [12] moving tries to move the output Threshold toward low cost classes such that examples with higher costs become harder to be misclassified. Threshold-moving is a good Many areas are affected by class imbalance problems. The solution provided by many techniques in data mining is helpful but not enough. The consideration of which technique is best for handling a problem of data distribution [13] is highly depends upon the nature of data used for experiment.

I.SIMPLIFIED SWARM OPT MIZATION (SSO)

SSO is a simplified version of Partial swarm optimization and can be used to find the global minimum of nonlinear Functions. In every generation, the particle's position value in each dimension will be kept or be updated by its p best value or by the gbest value or be replaced by new random value according to the procedure depicted in equation (1) Simplified swarm optimization technique detect he data [8]
may be redundant, noisy or irrelevant in nature. and reduce false positive rate.

$$x_{nd}^{t}= \begin{cases} x_{nd}^{t-1} & if \quad rand() \in [0.\, c_w) \\ p_{nd}^{t-1} & if \quad rand() \in [c_w,\, c_p) \\ g_{nd}^{t-1} & if \quad rand() \in [c_p,\, c_g) \\ x & if \quad rand() \in [c_g,\, 1) \end{cases}$$

$C_W < C_p < C_g$---------- (1)

Where i = 1; 2; m, where m is the swarm population. Xi

position of particles. Cw, Cp and Cg are three prede-termined positive constants with Cw < Cp < Cg. Pi = (pi1; pi2 ; : : : ; piD) denotes personal best and Gi = (gi1; gi2 ; : : : ; giD) denoted global best solution. The x represent new particle with random value between 0 and 1. The SSO algorithm is used to optimize membership function. The fitness function has been evaluated for each individual based on detection rate and false alarm rate to increase accuracy and efficiency of the proposed system.

I. SSO-ADABOOST-KNN ALGORITHM

In this algorithm [14] , Initially, the number of swarm population size, the number of maximum generation, and three parameters are determined. After initialization, SSO is conducted to search for best subset from both discrete and continuous variables of all features in dataset. In each iteration of SSO, Adaboost-KNN [15] is employed to classify

The subset of features selected by each particle. It filters data and reduce irrelevant.When training the Adaboost-KNN classifier [16], each KNN firstly returns a probability matrix Knn_score, then after T times training, Ada_score is computed according to all the T Knn_scores. Finally we computeThis approach is significantly different from other research work which had combine only data mining and PSO. The proposed method produce high efficiency and produce near optimal solution for pre-processing phase.

Algorithm

Step 1: Initialize the swarm size (m), the maximum generation (maxGen), the maximum fitness value (maxFit), Cw, Cp and Cg.

Step 2: In every iteration, a random number R that is in the range of 0 and 1 will be randomly generated for each dimension.

Step 3: Perform the comparison strategy where:
if ($0 \leq R < Cw$), then {xid = xid};
Else if ($Cw \leq R < Cp$), then {xid = pid};
Else if ($Cp \leq R < Cg$), then {xid = gid};
Else if ($Cg \leq R \leq 1$), then {xid = new (xid)};

Step 4: This process will be repeated until the termination condition is satisfied.

Step 5:nitialize distribution: D1(p)<-1/M (M is the size of training set)

Step 6: For t=1 to T
Step 7: Train KNN classifier under the distribution Dt and get the hypothesis ht : datat->c1; c2; …ckg,

compute probability matrix Knn_scoret 12

Step 8: Calculate the error item for ht
Output: Best feature subset sbest, classification result.

III. Experimental Result

We use the non-relational Glass and Shuttle datasets obtained from the UCI machine learning database [17]. We have chosen these datasets because they represent highly imbalanced datasets with different numbers of pattern classes which was treated as the positive class, and another represents the normal category which was treated as the negative class. These data sets were selected as the class imbalance problem, inherent in the data, hindered the learning from building an effective classification model.
Accuracy: Accuracies of each approach calculated using this equation:

Accuracy= TN+TP/(TP+TN+FN+FP)

Where TP, TN, FP, and FN stand for true positive, true negative, false positive and false negative, respectively.

Error rate: Error rate is the total number of incorrectly classified instances among all web pages available during the test.

1. Accuracy Comparison

The proposed SSO-Adaboost-KNN classification algorithm [20] achieves higher classification accuracy of 96.37%, which is 2.125 % and 3.408 % higher than the existing classification algorithm such as SSO-KNN and Adaboost-KNN classification

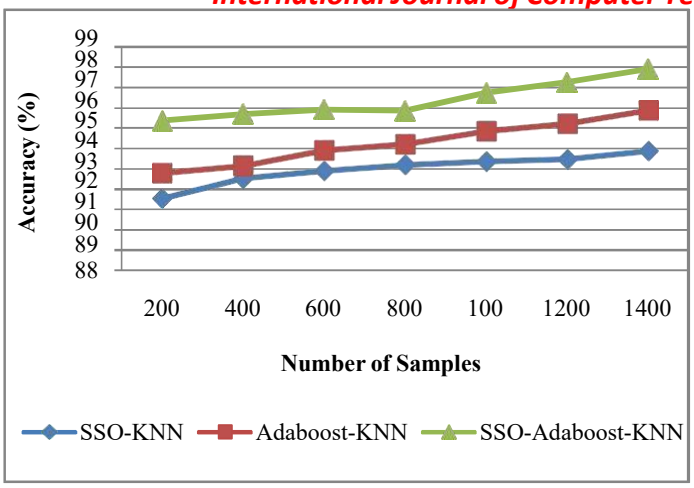

Fig 1: Classification Accuracy Comparison

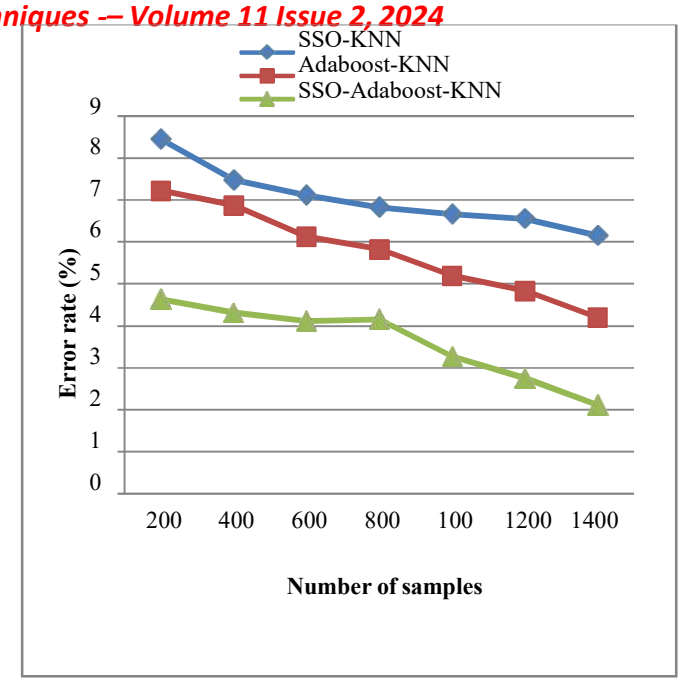

Fig 2: Error Rate Comparison

2. Error Rate Comparison

The classification error rates on the number of samples. From the graph the proposed SSO-Adaboost-KNN algorithm yields overall lower error rates of 3.62 %, which is 2.125% and 3.408% lower error rate when compared to SSO-KNN and Adaboost-KNN algorithms [18] respectively. The reason is that, the proposed work uses an SSO-Adaboost-KNN algorithm for data classification, where a set comprising of concurrent, distributed agents together find a meaningful organization of objects for a particular dataset. Also, the classification approach is highly effective than then the existing methods such asSSO-KNN and Adaboost-KNN classification.

IV. CONCLUSION

In this paper reported that data preprocessing provide Better solution than other methods because it allow adding new information or deleting the redundant information, which helps to imbalanced the data. Another method that helpfulto solve the problem of class imbalance is boosting. Boosting is powerful ensemble learning algorithm that improved the performance of weak classifier. The algorithm such as AdaBoost, SSO,KNN-AdaBoost is an example of boosting algorithm. Feature selection method can also used for classification [19] of imbalance data. The performance of a feature selection algorithm depends on the nature of the problem. Finally, this paper suggests that applying two or more technique i.e. hybrid approach gives better solution for class imbalance problem.

V. REFERENCES

[1] A. Bland and G. P. Dawson, “Tabu Search and Design Optimization”, Vol. 23, No. 3, pp. 195-201, April 1991.

[2] .Ravi, Chinapaga, Et Al. "Analysis Of Concept Drift Detection–A Framework For Categorical Time Evolving Data."

[3] Bhushan, P. Vinay, Et Al. "An Efficient System For Heart Risk Detection Using Associative Classification And Genetic Algorithms."

[4] F. Glover, “ A User’s Guide To Tabu Search “, Annals of Oper. Reas. 41(1993) 3-28

[5] Chnarsimha Chary,International Journal Of Scientificresearch In Computer Science, Engineering And Information Technology, “Duo Mining Techniques In Knowledge Discovery Process In Data Base”, 2018/06, Volume 3, Issue 1

[6] Bhushan, P. V., Nitesh, V., Chary, C. N., & Gupta, K. G. Novel Approach For Multi Cancers Prediction System Using Various Data Mining Techniques.

[7] Gupta, K. Gurnadha, Ch Narasimha Chary, And A. Krishna. "Study On Health Care Life Log By The Level Of Care Required Using Keygraph Technology In Text Data Mining."

[8] F. Glover, “Artificial Intelligence, Heuristic Frameworks and Tabu Search”, Managerial and Decision Economics, Vol.11 ,365- 375(1990).

[9] chnarasimha chary, international journal of research,, “classification of machine learning techniques and applications in artificial intelligence”, 2019/02,volume 1, issue 1.

[10] Narasimhacharycholleti, “analyzing security of biomedical data in cancer disease”, journal of critical reviews,2020/5, volume 7, issue 7, pages 150-156

[11] cholleti, narasimhachary, and tryambak hirwarkar."biomedical data analysis in predicting and identification cancer disease using duo-mining." advances in mathematics: scientific journal 9 (2020): 3487-3495

[12] chnarasimha chary,journal of critical review,”analyzing security of biomedical data in cancer disease”,volume 9, issue 1.

[13] cholleti, narasimhachary, and tryambak hirwarkar. "biomedical data analysis in predicting and identification cancer disease using duo-mining." *advances in mathematics: scientific journal* 9 (2020): 3487-3495.

[14] dr.narasimha chary ch, the international journal of analytical and experimental modal analysis ,“privacy preserving media sharing with scalable access control and secure deduplication in mobile cloud computing”, 2023/1, volume 15,issue 1, pages 150-156

[15] Ch, Dr. "Narasimha Chary,”." Comprehensive Study On Multi-Operator Base Stations Cell Binary And Multi-Class Models Using Azure Machine Learning”," A Journal Of Composition Theory 14.6 (2021).

[16] dr.ch.narasimha chary, a journal of composition theory,comprehensive study on multi-operator base stations cell binary and multi-class models using azure machine learning.

[17] Reddy, S. K., Radhika, G., Bharathi, K. S., & Chary, C. N. Comprehensive Study on Multi-Operator Base Stations Cell Deployment of B5G Utilizing Blockchain-Enabled SDN Architecture.

[18] Dr.Narasimha Chary Ch,The International journal of analytical and experimental modal analysis, ”Privacy Preserving Media Sharing With Scalable Access Control And Secure Deduplication In Mobile Cloud Computing”,,2023/1,Volume 15, Issue 1, Pages 150-156.

[19] F. Glover, “Tabu Search-Part I”, ORSA Journal on Computing, Vol. l., No. 3, pp. 190-206, 1989.

[20] CH, Narasimha Chary, et al. "Big Data in Healthcare Systems and Research ."

ISSN: 2394-2231 http://www.ijctjournal.org