International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume 6, Issue.3 E-ISSN: 2347-2693

A Novel Algorithm for Class Imbalance Learning on Big Data using Uniform Sampling Strategy (USS) Technique

Mohammad Imran

Dept. of Computer Science and Engineering, Muffakham Jah College of Engineering and technology,
(Affiliated to Osmania University, Hyderabad, Telangana)
Banjara Hills, Hyderabad, Telangana, India

Author's Mail Id: imran.quba@gmail.com

Available online at: www.ijcseonline.org

Received: 18/Feb/2018, Revised: 22/Feb2018, Accepted: 25/Mar/2018, Published: 30/Mar/2018

Abstract— Big data consists of large volumes of data which are used to discover the hidden knowledge. Class imbalance nature is a conventional issue which is present in all real world datasets. The class imbalance nature in the big data reduces the performance of the existing classification algorithms. The data source of diverse nature available from varied sources also degrades the performance of the existing algorithms. To address these issues of class imbalance problem the present work proposed various novel and effective class imbalance learning (CIL) algorithms. In this work, we proposed Uniform Strategic Sampling (USS) Technique novel algorithms approaches for class imbalance data sources.

Keywords—Class Imbalance Learning(CIL),Big Data,Sampling,Uniform Sampling Strategy Technique,Classification

I. INTRODUCTION

Decision trees are the mathematical based algorithmic model which uses logic as the core unit for decision making. Decision tree consists of the branches and leaves. Each branch is a path of splitting the records in to a narrow space and each leaf is the result of the classification of records in a specific class. There are numerous models of decision trees, which access the data and classify them in the predefined classes.

II. RELATED WORK

Rukshan Batuwita et al., [1] have studied the SVMs models on imbalance data learning and concluded that the learning process tends to improve the majority class and decreases the predictive ability for minority class. Rushi Longadge et al., [2] have gathered the evidence to show that a large number of existing algorithms build model to better predict majority class examples due to availability of examples and mistakenly classifies minority instances in to wrong classes when imbalance dataset are applied. Kun Jiang et al., [3] have developed a hybrid algorithm GASMOTE using genetic algorithm for resample of instances in the SMOTE approach and they also used an optimal threshold for minority sampling guided by genetic algorithm. Shaza M. Abd El rahman et al., [4] have reviewed the latest trends in the field of class imbalance learning, which provided novel solutions to the concern issues. Bartosz Krawczyk [5] has provided a study for varied benchmark solutions for different fields in the data mining like classification, clustering, uncertainty stream learning and data with large volumes and complexity. The assessment of the current works suggests that the efficiency of the decision tree reduces drastically when applied for class imbalance data sources. The reason for the reduce in performance is due to the inefficient model built with the rare instances class.

III.METHODOLOGY

III.I). PROPOSED UNIFORM SAMPLING STRATEGY (USS) FRAMEWORK

II). Components of Uniform Sampling Strategy (USS)

This sub section details uniform sampling method and its chief properties will be provided below as follows. Key components of the proposed approach are explained in phases.

In the first phase, the class biased data source is divided as majority N and minority P subspaces. However, the user proposed novel algorithm is a under sampling procedure, we require to center our attention on the majority sub space. In the next step of the algorithm, the ratio of examples needed in the majority subset for forming uniform subset is determined by the number of instances in minority subset. The amount of under sampling in the majority subset will be subject upon the unique characteristics of the data source.

After eliminating surplus instances randomly a new majority subset Ni is formed. The subclass with more number of instances Ni and subclass with less number of instances P are united which are combined to make a possible unbiased data source. The formed data source which is almost unbiased is executed on traditional procedure; according to the particular situation C4.5 was implemented for attaining diverse metrics of Area Under Curve, Recall, sensitivity, FP Rate etc.

IV. RESULTS AND DISCUSSION

IV.I Experimental Results

In this sub section, we perform the practical assessment of user proposed approach with the standard algorithms. The author's goal is to evaluate quite a few questions about the proposed investigational points to verify in the proposed approach under different scenario of binary biased class scenario.

1) In the initial scenario, authors wish for investigate regarding to establish superiority of any methods which is capable of managing the excess class biased nature in the data sources of diverse Imbalance Ratios i.e., to represent the utmost superior and efficient technique.
2) The authors wish to examine the enhancement in relation traditional classification approaches so as to examine suitability in the usage with the existence of an exclusive intermediate process level approach. The authors aimed to investigate the ratio among complication increase and efficiency improvement is acceptable or not.

In this context, the evaluation of the novel approach to be compared with all the existing approaches separately. This method presents the explained analysis for gaining a superior idea with required outcome by recognizing the advantages or restrictions of user proposed approach with all the considered approaches.

Hepatitis Dataset: The majority and minority ratio of the dataset is very high (i'e.123:32). The results of the tenfold cross validation are shown in Table 4. From table 4 we can conclude that proposed algorithm has given good results on all the measures.

Ionosphere Dataset: The majority and minority ratio of the dataset is moderately high (i'e: 225:126). From Table 5, we can observe the results of proposed algorithm Vs various algorithms diverse metrics of Area Under Curve, Recall, sensitivity, FP Rate etc. From the table we can conclude that proposed algorithm has given moderate results on Breast-w dataset.

Labor Dataset: The majority and minority ratio of the dataset is moderately high (i'e: 37:20). From Tables 6, we can observe the results of proposed algorithm Vs various algorithms diverse metrics of Area Under Curve, Recall, sensitivity, FP Rate etc. From the table we can conclude that proposed algorithm has given good results on Labor dataset.

Table 4.1 Summary of tenfold cross validation performance for Hepatitis dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.751±0.211	0.765±0.251	0.701±0.217	0.702±0.268	0.765±0.255
C4.5	0.668±0.184●	0.510±0.371●	0.409±0.272●	0.374±0.256●	0.900±0.097○
CART	0.563±0.126●	0.232±0.334●	0.179±0.235●	0.169±0.236●	0.928±0.094○
REP	0.619±0.149●	0.293±0.386●	0.210±0.259●	0.187±0.239●	0.942±0.093○
SMOTE	0.792±0.112○	0.709±0.165●	0.677±0.138●	0.681±0.188●	0.837±0.109○

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Table 4.2 Summary of tenfold cross validation performance for Ionosphere dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.913±0.065	0.929±0.069	0.896±0.067	0.874±0.102	0.928±0.074
C4.5	0.891±0.060●	0.895±0.084●	0.850±0.066●	0.821±0.107●	0.940±0.055○
CART	0.896±0.059●	0.868±0.096●	0.841±0.070●	0.803±0.112●	0.921±0.066●
REP	0.902±0.054●	0.886±0.092●	0.848±0.067●	0.826±0.104●	0.933±0.063○
SMOTE	0.904±0.053●	0.934±0.049○	0.905±0.048○	0.881±0.071○	0.928±0.057○

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Table 4.3 Summary of tenfold cross validation performance for Labor dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.913±0.154	0.918±0.252	0.818±0.262	0.775±0.305	0.968±0.155
C4.5	0.726±0.224●	0.696±0.359●	0.636±0.312●	0.640±0.349●	0.833±0.127●
CART	0.750±0.248●	0.715±0.355●	0.660±0.316●	0.665±0.359●	0.871±0.151●
REP	0.767±0.232●	0.698±0.346●	0.650±0.299●	0.665±0.334●	0.765±0.194
SMOTE	0.833±0.127●	0.871±0.151●	0.793±0.132●	0.765±0.194●	0.847±0.187●

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Table 4.4 Summary of tenfold cross validation performance for Breast-w dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.950±0.039	0.953±0.046	0.943±0.031	0.936±0.049	0.952±0.050
C4.5	0.957±0.034○	0.965±0.026○	0.962±0.021○	0.959±0.033○	0.932±0.052●

CART	0.950±0.032 0.968±0.026○ 0.959±0.020○ 0.952±0.034○ 0.940±0.051●
REP	0.957±0.030○ 0.965±0.030○ 0.960±0.021○ 0.957±0.033○ 0.931±0.060●
SMOTE	0.967±0.025○ 0.974±0.024○ 0.960±0.022○ 0.947±0.035○ 0.975±0.024○

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Table 4.5 Summary of tenfold cross validation performance for Colic dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.820±0.079	0.782±0.088	0.812±0.070	0.857±0.102 0.751±0.129
C4.5	0.843±0.070○	0.851±0.051○	0.888±0.044○	0.931±0.053○ 0.717±0.119●
CART	0.847±0.070○	0.853±0.053○	0.890±0.040○	0.932±0.050○ 0.720±0.114●
REP	0.844±0.067○	0.857±0.056○	0.882±0.043○	0.914±0.066○ 0.731±0.121●
SMOTE	0.908±0.040○	0.853±0.057○	0.880±0.042○	0.913±0.058○ 0.862±0.063○

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Table 4.6 Summary of tenfold cross validation performance for Pima Diabetes dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.753±0.065	0.730±0.065	0.733±0.058	0.743±0.093 0.716±0.096
C4.5	0.751±0.070●	0.797±0.045○	0.806±0.044○	0.821±0.073○ 0.603±0.111●
CART	0.743±0.071●	0.782±0.042○	0.812±0.040○	0.848±0.066○ 0.554±0.113●
REP	0.754±0.060●	0.785±0.037○	0.809±0.037○	0.838±0.072○ 0.567±0.105●
SMOTE	0.791±0.041○	0.781±0.064○	0.741±0.046○	0.712±0.076● 0.807±0.077○

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Table 4.7 Summary of tenfold cross validation performance for Vote dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.978±0.030	0.979±0.034	0.975±0.030	0.972±0.047 0.978±0.035
C4.5	0.979±0.025	0.971±0.027●	0.972±0.021●	0.974±0.029○ 0.953±0.045●
CART	0.973±0.027●	0.971±0.028●	0.966±0.022●	0.961±0.037● 0.953±0.046●
REP	0.957±0.023●	0.969±0.035●	0.961±0.025●	0.955±0.034● 0.949±0.059●
SMOTE	0.984±0.017○	0.977±0.027●	0.969±0.021●	0.963±0.037● 0.981±0.023○

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Table 4.8 Summary of tenfold cross validation performance for Sonar dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.805±0.088	0.819±0.101	0.786±0.099	0.771±0.140 0.822±0.114
C4.5	0.753±0.113●	0.728±0.121●	0.716±0.105●	0.721±0.140● 0.749±0.134●
CART	0.721±0.106●	0.709±0.118●	0.672±0.106●	0.652±0.137● 0.756±0.121●
REP	0.746±0.106●	0.733±0.134●	0.689±0.136●	0.685±0.192● 0.762±0.145●
SMOTE	0.814±0.090○	0.863±0.068○	0.861±0.061○	0.865±0.090○ 0.752±0.113○

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Table 2.9 Summary of tenfold cross validation performance for Sick dataset

System	AUC Summary	F-measure	TP Rate	TN Rate
USS	0.805±0.088	0.972±0.036	0.969±0.026	0.968±0.033 0.968±0.041
C4.5	0.726±0.224●	0.696±0.359●	0.636±0.312●	0.640±0.349● 0.833±0.127●
CART	0.750±0.248●	0.715±0.355●	0.660±0.316●	0.665±0.359● 0.871±0.151●
REP	0.767±0.232●	0.698±0.346●	0.650±0.299●	0.665±0.334● 0.765±0.194●
SMOTE	0.833±0.127○	0.871±0.151●	0.793±0.132●	0.765±0.194● 0.847±0.187●

○ Empty dot indicates the loss of USS.
● Bold dot indicates the win of USS;

Breast-w Dataset: The majority and minority ratio of the dataset is moderately high (i'e: 458:241). From Table 7, we can observe the results of proposed algorithm vs various algorithms diverse metrics of Area Under Curve, Recall, sensitivity, FP Rate etc. From all the tables we can conclude that proposed algorithm has given moderate results on Breast-w dataset.

Diabetes Dataset: The majority and minority ratio of the dataset is very high (i.e. 500:268). From Table 9, we can observe the results of proposed algorithm Vs various algorithms diverse metrics of Area Under Curve, Recall, sensitivity, FP Rate etc. From all the tables we can conclude that proposed algorithm has given good results on AUC and tie and some underperforming results in the case of remaining measures.

Vote Dataset: The dataset majority and minority ratio is moderately high (i'e: 287:168). From Table 10, we can observe the results of proposed algorithm Vs various algorithms diverse metrics of Area Under Curve, Recall, sensitivity, FP Rate etc. From all the tables we can conclude that proposed algorithm has given moderate results on Breast-w dataset.

Sonar Dataset: The multi class nature and the majority and minority ratio of the dataset is moderately high (i'e: 111:97). From Table 11, we can observe the results of proposed algorithm vs various algorithms diverse metrics of Area Under Curve, Recall, sensitivity, FP Rate etc. From all the tables we can conclude that proposed algorithm has given good results on Sonar dataset.

Sick Dataset: The dataset and the majority and minority ratio of the dataset is moderately high (i'e: 3541:231). From Table 12, we can examine the consequences of proposed approach verses various compared approaches diverse metrics of Area Under Curve, Recall, sensitivity, FP Rate etc. From all the given tables we can come to a point that proposed approach has given outstanding results on Sick dataset.

To conclude, we can say that the proposed approach is one of the finest alternatives to handle class biased problems efficiently. This investigational study claims the findings that the uniform sampling of both majority and minority subset can get better with the CIL behaviour when handling with class biased data sources, as it has improved the proposed approach to be the best executing approach when evaluated with four classical and well-known approaches: C4.5, CART, REP and SMOTE a well-established under sampling algorithm.

V. CONCLUSION AND FUTURE SCOPE

This section provides a novel approach on the problem of biased class distributed data. This approach implements, unique uniform sampling strategy. The data source is approximately balanced in a way to improve overall metrics for supervised learning. The experimental outcomes suggest that the proposed approach noticeably decreases imbalance scenario at same time preserving and increasing classification metrics with the existing approaches. However, the new algorithm is very much helpful for constructing of improvised decision trees on different datasets to generate varied and improved validation metrics.

REFERENCES

[1]. Rukshan Batuwita and Vasile Palade, "CLASS IMBALANCE LEARNING METHODS FOR SUPPORT VECTOR MACHINES", Imbalanced Learning: Foundations, Algorithms, and Applications, By Haibo He and Yunqian Ma, Copyright c **2012** John Wiley & Sons, Inc.

[2]. Rushi Longadge, Snehlata S. Dongre, Latesh Malik," Class Imbalance Problem in Data Mining: Review", International Journal of Computer Science and Network (IJCSN) **Volume 2, Issue 1, February 2013.** www.ijcsn.org ISSN 2277-5420.

[3]. Kun Jiang, Jing Lu, Kuiliang Xia," A Novel Algorithm for Imbalance Data Classification Based on Genetic Algorithm Improved SMOTE", Arab J Sci. Eng, DOI 10.1007/s13369-016-2179-2.

[4]. Shaza M. Abd Elrahman and Ajith Abraham, "A Review of Class Imbalance Problem" Journal of Network and Innovative Computing ISSN 2160-2174, **Volume 1, pp. 332-340, 2013.** ©MIR Labs, www.mirlabs.net/jnic/index.html

[5]. Bartosz Krawczyk," Learning from imbalanced data: open challenges and future directions", Prog Artif Intell, DOI.10.1007/s13748-016-0094-0.

AUTHOR PROFILE

Mohammad Imran received his B.Tech (CSE) in 2006 and M.Tech (CSE) in 2008 from JNTU, Hyderabad both with distinction , His Research interests include, Deep learning, Artificial Intelligence, Neural Networks, Class Imbalance Learning, Ensemble learning, Machine Learning, Data mining and Big Data Analytics. He is Pursuing Doctor of Philosophy [Ph.D (CSE)] in the Department of Computer Science and Engineering, Rayalaseema State Government University, Kurnool-518007, Andhra Pradesh.

He has published more than 12 research papers in reputed international journals including Scopus Indexed SCI & Web of Science and conferences including IEEE and it's also available online.

He is the recipient of **Pradhan Mantri Kaushal Vikas Yojna (PMKVY) Certified Telecom Terminal Equipment Application Developer** in **November 2017** Conducted by **Telecom Sector Skill Council Development under National Skill Development Corporation (NSDC)**.He has taught more than 20 PG and UG Subjects including Artificial Intelligence, Neural Networks, Machine Learning, Data Warehousing & Mining, Information Retrieval Systems, Database Management Systems, Distributed Databases,, Object Oriented Analysis & Design, Software Testing Methodologies, Object Oriented Programming, Data Structures, Operating Systems, Mathematical Foundations for Computer Science, Computer Graphics, Computer Networks, Programming Languages like Python, JAVA, C++, C and Scripting Languages He is a Member of IEEE,Computer Society of India and ACM.

