Methods to Handle Multiclass Imbalance Data in Educational Data Mining

Bhasha Anjaria, Ankita Gandhi, Jay Gandhi

Abstract: In Scientists ordinarily exclude the equalization of the dissemination on a dataset in Educational Data Mining (EDM). It can truly influence the consequence of the classification procedure. Hypothetically, the distribution of data is respectively balanced pretended by the majority of classifier. Hence, the execution of the classification algorithm simply turned out to be less viable and should be taken care of the issue could illuminated. These exploration would characterize about imbalanced class on multiclass EDM dataset minding component utilizing the Map Reduce. This strategy serves adjusting system for the dataset's dissemination, using parallel processing; those classification result will the results. These balancing strategies can be implemented with different kind of classification methods like Naïve Bayes, SVM, NN to measure the improvisation in the results.

Keywords : Educational Datamining, Imbalanced class classification, MapReduce, Multiclass, Resampling Techniques .

I. INTRODUCTION

Data mining is a flexible field of concentrate that connect the AI, design acknowledgment, insights, data base, perception on data base data evocation's issues [1]. It is an utilized on a few sorts of field like therapeutic, designing, economy, instruction and so forth. In information preparing method inspecting are connected on information in which either new example are included or existing examples is evacuated. Technique of summing latest samples in old samples are noted as over-sampling & remove sample technique is known as under sampling. Data mining are the famous approach to evaluate understudies achievement. Presently a day's data mining have connected in teaching region and it is called educational data mining.

Information examination on instructive region is for the most important part that can be known as Educational Data Mining. One of the uses of EDM are survey & gathering understudy's data on studying method and knowing conclusion [2].

Revised Manuscript Received on April 02, 2020.

Bhasha Anjaria, lecturer, Information Technology, Parul Polytechnic Instituute, Vadodara, Gujarat

Ankita Gandhi, Deputy HOD & an Assistant Professor in Computer Science and Engineering Department,PIET, Parul University Vadodara, Gujarat, India

Jay Gandhi , Assistant Professor in Computer Science and Engineering Department,PIET, Parul University Vadodara, Gujarat, India

EDM is bothered to create , inquire and apply automated techniques to identify designs in huge gathering of instructive information that would some way or in another way of difficult assessment because of the tremendous volume of information inside which is prevailed. There are packs of procedures that can be used in EDM issues, for instance, relapse, order, and classification. Order systems are a most loved strategies to assess and group the understudy's execution [3]. In light of the past examinations, there are two distinctive ways that can be executed on the data level and on the algorithmic measurement. Information exact strategy is ordinarily done the pre-dealing with endeavor by modify or adjust the tendency of the class transport on the dataset.
consolidate is a champion among the most all around used system that used for the information level strategy.

II. RELATED WORK

Yoga Pristyanto et al. using Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification(2018), In this paper will explain about imbalanced class on multiclass EDM dataset handling mechanism using the combination of SMOTE and OSS. SMOTE and OSS method provides balancing mechanism for the dataset's distribution, so that the classification results will be enhanced in terms of classification performance. The result shows that the combination of SMOTE and OSS can enhance the performance of SVM as the classification method that used in this study. Those combination of methods produce the accuracy, sensitivity, specificity, and g-mean score as high as 88.637%, 92.292%, 95.554%, 93.796% respectively. For further research it is hoped that research can be done especially for testing the proposed methods using other classification algorithms such as k-NN, Naïve Bayes, and Decision Tree.[1]

Mohammad Imra, et al. Using Weka Tool is using in Information Mining of Imbalanced Dataset in Educational Data (2016). In this paper data mining approach has been utilized in trade target from it is beginning notwithstanding, recently it is utilized effectively for new & developing territories like training structure. In these exploration, we utilize data mining ways deal with anticipate understudies' last result, i.e., last result in specific system by defeating issue of imbalanced dataset. I execute a few re-inspecting strategies adjust dataset so that can improve execution. Re-inspecting systems incorporate Synthetic Minority Over-testing Technique, Random over Sampling.

Trial results demonstrate that re-testing procedures improve the execution of order model that is produced to anticipate understudies' last result in a specific area. For future work, this study will be useful for establishments and commercial ventures. We can be producing the data in the wake of actualizing the others information mining systems like bunching, Predication and Association rules and so forth with help of Data Mining devices.[2]

SyedTanveerJishan, et al. utilizing improving precision of understudies' last grade to use perfect comparable width binning and engineered minority over-inspecting strategy. (2015) in this paper Educational data mining is a starting late discernible area in the field of data mining and it may be related with better understanding the enlightening structures. In this paper, we are accessible how data can be pre-managed with the utilization of discretization methodology called the Optimization of Equal Width Binning and an over-testing system known as the SMOT to improve the accuracy of the understudies' last category desire appear for a individual path. The result got from the preliminary gives an indisputable sign that the precision of the desire show improves on a very basic level when the discretization and over-looking at systems are associated. In future, we may in like manner need to examine how a comparative improvement framework capacities for other data binning methodologies like binning by repeat, binning by size .[3]

Raisul Rashu, et al. usage of Information Mining Approaches to imagine last Evaluation by Overcoming Class Imbalance Problem.(2017) In this investigation, we use information mining ways to deal with oversee foresee understudies' definitive result,, i.e., In the last category specific course by beating the issue of uneven dataset. We complete a couple of re-testing techniques to change the dataset so that could indicate advance execution. Reassessment of systems to join SMOT, i. Random under Sampling, II Random over Sampling. Test outcomes exhibit that reevaluating strategies redesign the execution of the request models that are made to anticipate understudies' last grade in a particular course. The investigate approach is isolated into II phases. Phase 1 is to develop a model that will predict the dimension of the understudies for a particular course. Looking the execution in stage 1, organize 2 Further improvement needed to anticipate the class anomaly problem.[4]

Mr.RushiLongadge, et al. utilizing Class Imbalance Problem in Data Mining: Review(2013) For dispensing perfect loads for the determined highlights we propose a novel strategy in this exploration. This proposed methodology is attempted on CE CT Lung pictures. Reproduction results and examination gave the idea that our proposed structure has shown better grouping exactness of tumor types than the normal SVM. Dealing with the nonlinear information. This strategy is Easy to create principles and Easy to get it. Further this strategy used for separation among amiable and dangerous tumors. More highlights can be removed and allotted utilizing this strategy for better division and classification.[5]

III. DIFFERENT METODOLOGIES

A. Data Acquisition

This procedure is readied the information that are utilized on this examination. In this set information is caught from the open informational index. The informational index is paired information with 403 occurrences. Each example contains 4 qualities and 1 mark or class. Quality are demonstrates the normal for information, and mark class focus in information. in these information, four mark is lowest with fifty cases, low 129 occurrences, center 122 examples, and high 130 cases.

B. Data pre-processing

In information pre-processing the data will be cleaned, changed and arranged to be prepared. In this pre-processing method imbalance data distribution will be done. In this study under sampling and over sampling are used. OSS used as a under sampling technique and Synthetic minority over-sampling technique used as a over sampling technique.

Synthetic minority over-sampling technique (SMOTE)

On genuine data additional preparation data by playing out specific tasks created by them. for that scenario, tasks like rotation and skew the preparation data were regular approaches to irritate by activities like rotation and skew. The minority class is over-investigated by taking every minority class test and showing built perspectives along the line fragments joining any/the vast majority of the k minority class closest neighbors. Subordinate upon the extent of over-analyzing required, neighbors from the k closest neighbors are discretionarily picked.

OSS (One sided selection)

OSS method is divided by Four section : Borderline pattern, Safety pattern, Noise pattern, and Redundant pattern.
Noise pattern: The greater part class is encompassed by minority class.
Borderline pattern: The majority class located in the middle between the two classes.
Redundant pattern: The dominant part class that is situated a long way from as far as possible.
Safety pattern: The majority class have important information.

Map Reduce :

Map Reduce are programming structure those enables us to achieve conveyed and coordinate handling on extensive information indexes in disseminated domain. Map reduce consists comprises of two undertaking Map and Reduce As the label Map Reduce recommends, reducer stage happens after mapper stage has been finished.

Along these lines, the first is the map work, where a square of data is perused and prepared to create basic-esteem combines as middle of the road yields. Output of a Mapper or map job (key-value pairs) is input to the Reducer.

The reducer receives the key-value pair from multiple map jobs. The reducer aggregates those intermediate data tuples (intermediate key-value pair) into a smaller set for tupls or basic-esteem sets which are last yield.

C. Classıfıcatıon

SVM

SVM or Support Vector Machine is a linear model for classification & relapse issue. It can take care of straight and non-direct issues and function admirably for some down to practical issues. The possibility of SVM is straightforward: The algorithm makes a line or a hyperplane which isolates the data into classes.

RF

Random forest (RF) is a standout amongst the most generally utilized and powerful machine learning strategies which has demonstrated higher precision rate among recent machine learning algorithms. It is reasonable for preparing huge arrangement for data with an assurance of assessing the most suitable highlights required for classification. Random Forest is a gathering of tree-organized classifiers where each tree relies upon the estimations of irregular vector tested autonomously and the dissemination of all trees in the forest.

KNN

KNN will be a hearty should loud preparing information. In KNN classifier, closest neighbor determines the choice limit mainly. To 1NN we relegate every report of the population about its storeroom neighbor. For KNN we relegate every archive of the dominant part class of its k closest neighbors the place k may be a parameter. KNN is not thick, as hearty. Those arrangement choices of every test archive depends on the class of a absolute preparation document, which might be erroneously named alternately an ordinary.

IV. PROPOSED APPROACH

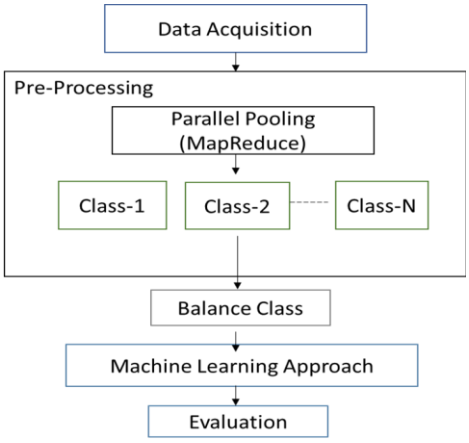

Fig. 1. Proposed Approach Block Diagram

In the proposed system consist four step namely preprocess, balanced and unbalanced class and machine learning approach/ classification and evolution. —Specialists generally overlook the adjustment of the course on a dataset In Educational Data Mining. It can truly influence consequence of classification procedure. Data mining is an versatile area for research those bring together machine learning, design acknowledgment, measurements, data base, and perception on data base data eradication's issue.

Algorıthm

Step 1: Data Acquisition
Step 2: MapReduce: On Parallel Pulling Profile
Step 3: Divided data into linear classes
for i=1:n
for k=1:m
if data(i,j)==some_condition
lin_dat=data(i,k)
else
non_dat=data(i,k)
end
end
end

Step 4: Divided data into balance and undalance class
X_bal=[x1,x2,x3,…xn] x_unbal=[y1,y2,y3,…..yn];
Step 5: Apply machine learning approach SVM, ANN or RF
fn_dat=[x1,x2,x3,….xn]; fn_lab=[lb1,lb2,lb3……lbn];
tn_dat=[tn1,tn2,…tnn];
train_dat=(fn_dat,fn_lab)
Test_dat=predict(tn_dat)
Step 6: Evaluate the system using Accuracy, Precession and Recall.

V. RESULT AND ANALYSIS

This work majorly focuses on handling imbalance data where the classification of the data is not balanced. Thus further the work can be done to achieve a result by applying different sampling techniques with various classification techniques. Majorly it is seen in the literature w, which has been reviewed that a parallel processing will help in this area. To achieve parallel processing of resampling technique, a MapReduce approach will be used in further implementation literature.

VI. CONCLUSION & FUTURE WORK

Data Preprocessing plays a vital role to work with imbalance class in the field of data mining. Inclusion of various sampling methods like SMOTE (Synthetic Minority Over-sampling Technique) and OSS (One Sided Selection) to the preprocessing stage will lead us towards the good results. As a future work, different classification techniques (SVM, KNN, RF) can be applied in a combination with sampling techniques. Moreover parallelism between sampling techniques can be achieved using different approaches like MapReduce.

REFERENCES

1. Yoga Pristyanto, Irfan Pratama, Anggit Ferdita Nugraha , " Data Level Approach for Imbalanced Class Handling on Educational Data Mining Multiclass Classification" International Conference on Information and Communications Technology (ICOIACT- 2018).
2. Cristobal Romero et al, Data Mining in Education (2013) 12-27.
3. Mr.Rushi Longadge et al, Class Imbalance Problem in Data Mining, (IJCSN 2013).
4. Gongzhu Hu et al, Classification of Wine Quality with Imbalanced Data, (IEEE 2016) 1712-1717.
5. Syed Tanveer Jishan et al, Improving accuracy of understudies' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique, (Decision Analytics 2015) 1-25.
6. Mohammad Imran et al, Data Mining of Imbalanced Data in Educational Data Using Weka Tool, (IJESC 2016) 7666-7669.
7. Amirah Mohamed Shahiria et al, A Review on Predicting Student's Performance using Data Mining Techniques, (TISIC 2015) 414-422.
8. Raisul Islam Rashu et al, Data Mining Approaches to Predict Final Grade by Overcoming Class Imbalance Problem, (ICCIT 2014) 14-19.
9. Mingyue Luo,Gang Liu , Distributed log information Processing with Map-Reduce , IEEE - 2010
10. Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer , SMOTE: Synthetic Minority Over-sampling Technique , Journal of Artificial Intelligence Research (2002)
11. M. Han, J., & Kamber, Data Mining: Concepts and Techniques Second, Second Edi., vol. 12. San Fransisco: Morgan Kauffman, 2006.

12. E. Ayers, R.Nugent, and N. Dean. “A Comparison of Student Skill Knowledge Estimates”, In International Conference On Educational Data Mining, Cordoba, Spain, pp.1-10, 2009.
13. A.K. Pal, & S.Pal. “Analysis and Mining of Educational Data for Predicting the Performance of Understudies”, International Journal of Electronics Communication and Computer Engineering, vol. 4, issue 5, pp.1560-1565,2013.
14. G. H. Nguyen, A. Bouzerdoum, S.L. Phung, “A supervised learning approach for imbalanced data sets”, pp.1-4, ICPR 2008.
15. J Dean, S Ghemawat,” MapReduce: Simplified data processing on large clusters, ” Communications of the ACM, 2008
16. Shuo Wang, Member, and Xin Yao, “Multiclass Imbalance Problems Potential IEEE Analysis and Solutions”, Transactions On Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, No. 4, August 2012.
17. Qiong Gu, Zhihua Cai, Li Zhu, and Bo Huang. Data mining on imbalanced data sets. In Proceedings of International Conference on Advanced Computer Theory and Engineering, pages 1020–1024. IEEE, 2008.
18. Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. Journal of Artificial Intelligence Research, 16:321–357, 2002.

AUTHORS PROFILE

Bhasha Anjaria works at Parul Polytechnic Instituute, Vadodara, Gujarat as a lecturer in Information Technology Department. She is pursuing M.Tech from Parul University. Her area of interest is Data mining & Machine learning. She has published 2papers in reputed journals and conferences. (email: bhashaanjaria@gmail.com)

Ankita Gandhi is working with Parul University Vadodara, Gujarat, India as a Deputy HOD & an Assistant Professor in Computer Science and Engineering Department,PIET. She is currently pursuing his PhD from Gujarat Technological University. Her area of interest is Machine learning, Algorithms and Data Mining. She has published more than 15 research paper in reputed Journals and Conferences. (E-mail--ankita.gandhi@paruluniversity.ac.in)

Jay Gandhi is working with Parul University Vadodara, Gujarat, India as an Assistant Professor in Computer Science and Engineering Department,PIET. He is currently pursuing his PhD from Nirma University. His area of interest is Data mining, Machine learning, and Opportunistic network. He has published more than 10 research paper in reputed Journals and Conferences. (e-mail: jaygandhi7591@gmail.com)