A Recapitulation of Imbalanced Data

Shaheen Layaq, B. Manjula

Abstract: ***In today's authentic universe almost all applications are imbalanced. Data imbalance is growing faster than ever before as many systems are interested in extracting knowledge from lake of data. Imbalance issue occurs because required data is very rare and using that rare data if classification is done we may lead to inaccurate result. In few sensitive imbalance cases like medical and finance if classification is done health and wealth both will get a huge lost. It is observed that big data and imbalance issue are having hand in hand relationship. So, imbalance data is gaining much importance in data science. It is predicted that by the year 2020, about 1.7MB of lake of information will be created every second by each device due to development in science and technology. Almost this lake of information generated will be imbalanced. So, in this paper we will define big data and imbalanced data, how there are related to each other, some of the reasons why imbalance data problems are occurring, various areas where imbalance issues is been effecting, current four machine learning methods for imbalanced data (Data based method, Algorithm based method, Cost sensitive method and ensemble methods), overall performance evaluation of imbalance methods are done using a comparison chart and interpreting achievements of imbalanced data using confusion matrix, Combined evaluation measures (G-means, F-Measure, Balanced Accuracy, Youden Index and Matthews's correlation coefficient) and Graphical performance evaluation using Receiver operating characteristic (ROC) curve and Area under the curve (AUC) and lastly, considering of result of various imbalance methods.***

Keywords: Big data, imbalanced data, machine learning, ensembles.

I. INTRODUCTION

Now a day's modern systems are generating lake amount of information and every one are interested in extracting knowledge from it. There is no particular structure for lake amount of information it may be in the form of structured, unstructured or semi structured. But most of the data is unstructured. This lake unstructured data is known as *big data*. In 2001, Gartner's Doug Laney first presented big data as three V's but, later many V's were added to big data which can be shown in Fig. 1.

Many systems in the world such as industry, commercial and bank are interested in extracting knowledge from the lake of data. From past two decades it is observed that data size is increasing tremendously. In year 2011 the size of digital data

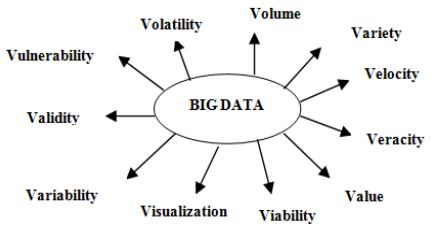

Fig. 1.The 11v's of big data

is roughly 1.8 Zettabytes and by the year 2020, the size will be 50 times more. And it is observed that almost big data generated is imbalanced which can be shown in Fig. 2.

So, if we consider big data as 'D' and imbalanced data as 'I' then we can denote it as I⊆D i.e. imbalanced data is part or almost equal to the big data which can be shown in the Fig. 3. Where ever big data is present imbalance data problem occurs. Imbalance data problem is seen in many operations like text mining [1], presence of oil spills detection [10], video mining [2], target detection [3], software defect prediction [4], sentiment analysis [5], industrial systems monitoring [6], cancer malignancy grading [7], behavior analysis [8], detection of fraudulent telephone calls [11], activity recognition [9] and so on. So, we can say where ever rare data is present imbalance data problem occurs. When there is lake amount of data and distribution of that data

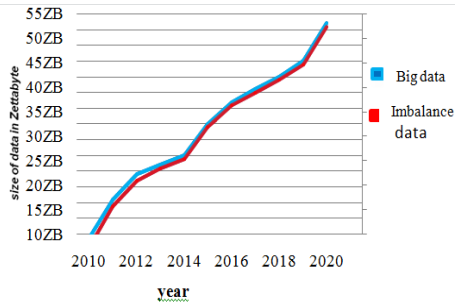

Fig. 2. Comparison of big data and imbalanced data

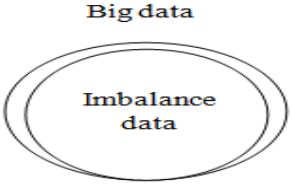

Fig. 3. Imbalance data is subset of big data

Among the classes is unequal then imbalance issue occurs. So, to solve this issue converting of imbalanced data to balance data is needed.

Revised Manuscript Received on January 05, 2020
***** Correspondence Author
Shaheen Layaq, Department of Computer Science, Kakatiya University, Warangal, Telangana, India.
Dr. B. Manjula, Department of Computer Science, Kakatiya University, Warangal, Telangana, India.

II. A LIGHT ON IMBALANCE DATA PROBLEM

If a class contains greater data count than the alternative classes then it is called imbalance dataset [12]. The greater data count class is known as huge or wrong or majority or negative class and less data count class is known as tiny or right or minority or positive class. Here right class is very rare and it is most important to us. So, this rare data leads to imbalance data problem. Imbalance data set can be shown in the Fig. 4.

If rare data problem occurs in finance and biomedical field it results in a great lost. For instance, in imbalanced cancer dataset right (minority) class represents patient who have cancer and wrong (majority) class represents patient who does not have cancer. Depending on wrong class if decision is taken patient may missed the best time for treatment. This is occurring because accuracy of wrong class is considered here. So, accuracy of right class has to be considered.

For classification of big data many machine learning approaches are developed such as logistic Regression, Decision trees, Neural Network & Naive Bayes. The class imbalance generally influences the performance of classical classification. The classical or traditional machine learning approaches doesn't work well with imbalanced data as there are dependent on general condition or common search but for imbalanced data specific search is required. Some of the reasons observed for imbalanced data problem are:

1) Required data may be in less number.
2) Due to missing values.
3) The sample picked from the 'D' dataset may contain fewer amount of required data.
4) Required data may present at extreme ends or at boundary line.
5) Due to incomplete data.
6) Due to large amount of redundant data.
7) Due to noisy data.
8) The required data may be considered as outlier.
9) Due to class overlapping.
10) Due to small disjuncts.

III. UNFOLDING IMBALANCE DATA PROBLEM

To unfold imbalanced data problem, imbalance data have to convert to balance data. For that from past two decades many research and studies were conducted.

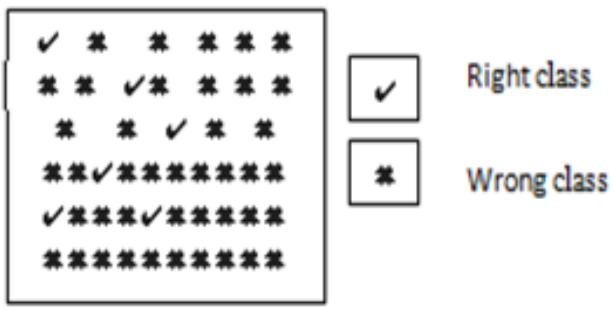

Fig. 4. Imbalance Dataset

Later many approaches have been proposed which basically falls into four categories. Data based, algorithm based, cost sensitive and ensemble method [13].

Data based also knows as external or sampling method. It targets on modifying the sample set by incrementing or decrementing samples to make suitable for standard learning algorithm. It is basically divided into three types oversampling, undersampling and hybrid sampling methods. *Oversampling* tries to highlight the right (minority) class by incrementing the data count of right class. Methods under oversampling are ROS, CBOS [20], SMOTE [14], B-SMOTE [15] and MWM[21]. On other hand *Undersampling* decrements the data count of wrong class so that right class is highlighted. RUS [16], CBUS [17], DSUS [18] and F-CBUS [19] are methods under it. *Hybrid sampling* is combination of oversampling and under sampling.

Algorithm based method [24] also knows as internal method. In algorithm method existing algorithms are updated for boosting the classification of the right class and performance of dataset. Decision tree, backpropogation neural network, Bayesian network, SVM, z-SVM, weighted nearest neighbor classifier, argument based rule learning and associative classification are some examples for it.

Distinct misclassification costs are allotted for distinct classes in *Cost sensitive method* [23]. Mainly it aims to assigns a higher weight for right class in order to make classifier give more priority to right class samples when it is used in imbalance data classification. It combines both algorithm and data level methods. C4.5, near Bayesian SVM, cost sensitive NN with PSO are few examples.

The *ensemble method* [22] are best for improving individual classifier achievements. It construct several two stage classifiers and the aggregate their predictions by considering original data. It contains compatible algorithms, cost sensitive operations and preprocessing. Bagging, boosting, boosting using cost sensitive and hybrid are some ensemble methods.

IV. INTERPRETING THE ACHIEVEMENTS OF IMBALANCED DATA

The classifiers performances are evaluated using evaluation matrices. The most popular evaluation technique is *confusion matrix* which is shown in Fig. 5. But, it works well with majority type of data. It is not good with imbalanced data.

	Guessed Positive	**Guessed Negative**
Real Positive	True Positive	False Negative
Real Negative	False Positive	True Positive

Fig. 5. Confusion Matrix

Table – I: Contrasting of various Imbalance Methods

Sno	Approaches	Methods		Merits	Demerits
1	*Data based Approach*	*Oversampling*	ROS,CBOS,SMOTE, BSMOTE, MWM.	• No information loss. • It performs well if sample is small.	• Perform worst if dataset is large. • Increases processing time. • Risk of over fitting.
		Undersampling	RUS,CBUS,DSUS, F-CBUS.	• Perform well if small amount of data is present • No need of trial and error process. • Improve run time and storage problems.	• Loss of data. • Performance worst if dataset is large because huge data is loosed.
		Hybrid	SMOTETomek, SMOTEENN.	It takes advantages of oversampling and undersampling	Both methods have to implement simultaneously.
2	*Algorithm Based Approach*	SVM ,Decision tree, Bayesian network, z-SVM, weighted nearest neighbor classifier, argument based rule learning , associative classification, backpropogation neural network..		• Depends only on internal . logic. • Effective algorithms because depending on problem algorithms are developed.	• Modification of existing learners. • It may need some preprocessing task.
3	*Cost Sensitive Approach*	Near Bayesian SVM, Cost Sensitive NN, PSO, C4.5, Cost Sensitive learning for SVM, SVM for Adaptively Asymmetrical Misclassification cost.		• It performs well if dataset is large. • It builds a model with minimum misclassification costs and process fastly.	• Problem when real cost value is unavailable (error cost is added when real cost is not known) . • Worst with small dataset.
4	*Ensemble Approach*	Bagging, Boosting, Booting using cost sensitive, hybrid.		• Simple, popular, easy to apply, competive , robust to difficult data , versatile and diverse method. • Used as esteemed method in solving imbalance problem.	• Hard to predict how much vast ensembles must built. • It depends on majority voting. • Complexity increases with increasing of methods.

Accuracy, error rate, sensitivity, specificity and precision metrics were added to confusion matrix to improve the performance when imbalanced data was considered which can be refer to "(1)".

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$
$$\text{ErrorRate=1-Accuracy=} \frac{FP+FN}{TP+TN+FP+FN}$$
$$\text{Sensitivity(OrRecall)=} \frac{TP}{TP+FN}\text{(Accuracy of positive examples)} \quad (1)$$
$$\text{Specificity=} \frac{TN}{TN+FP}\text{(Accuracy of negative examples)}$$
$$\text{Precision} = \frac{TP}{TP+FP}$$

Combined evaluation measures like G-means, F-Measure, Balanced Accuracy, Youden Index and Matthews's correlation coefficient refer to "(2)" combines both sensitivity and specificity by which more accuracy results are obtained in imbalanced dataset.

$$GM = \sqrt{Sensitivity\ X\ Specificity}$$
$$F_\beta = \frac{(1+\beta^2)\text{X Recall X } Precision}{\beta^2\text{X Recall X } Precision}$$
$$\text{BalancedAccuracy} = \frac{1}{2}(Sensitivity + Specificity) \quad (2)$$
$$\text{J= 2* BalancedAccuracy-1}$$
$$\text{MCC} = \frac{TPXTN-FPXFN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

Graphical performance evaluation is important as it is visualization metrics for evaluating any classifiers model performance. ROC (Receiver operating characteristic) and AUC (Area under the curve) curves are examples of it. ROC and AUC uses concepts of confusion matrices, specificity and sensitivity. ROC shows the vice versa relationship between true positives and false positives.AUC is used to know which model is best on average to evaluate the classifiers performance. Area of the graphic is considered to compute the AUC measure.

V. RESULTS

We study the performance of four approaches data based, algorithm based, cost sensitive and ensemble. Finally, now we can contrast the four imbalance approaches which can shown in Table--I. And to make comparison more accurate the overall performance can be shown with the help of a chart in Fig. 6. By viewing it we can understand that ensemble methods perform better than other methods.

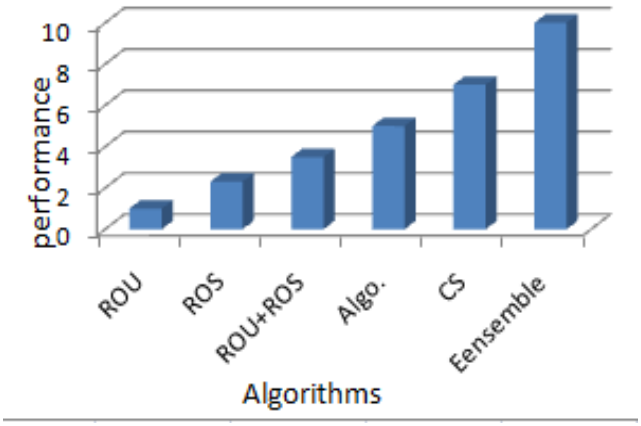

Fig. 6. Overall Performance of Imbalance Methods

VI. CONCLUSION

Here we tried to zoom on relationship between big data and imbalanced data, reason behind imbalance issue, areas effecting imbalance issue. And each machine learning method was considered in detail with their merits, demerits and performance.

After making all study we came to a conclusion that imbalanced data is a burning issue of big data which never comes to an end. To have accurate prediction model imbalance data have to convert to balance later, classification is to be done for that enhanced multiple methods have to consider. So, our forthcoming work will depend on advanced ensemble methods by which efficient and optimistic outcomes are obtained when imbalance issue occur.

REFERENCES

1. T.Munkhdalai, Oyun-Erdene Namsrai and K.H. Ryu," Self-training in significance space of support vectors for imbalanced biomedical event data," BMC Bioinformatics,Vol.16(Suppl 7):s6, April 2015.
2. Z.Goa,L.Zhang,M.-Y.Chen, A.G.Hauptmann,H.Zhang and A.-N.Cai,"Enhanced and hierarchical structure algorithm for data imbalance problem in semantic extraction under massive video dataset," Multimed. Tools Appl.,Vol. 68,Issue 3, 2014, pp. 641-657.
3. S.Razakarivony and F.Jurie," Vehicle detection in aerial imagery: a small target detection benchmark," Journal of visual communication and image representation, vol. 34, 2016, pp.187-203.
4. M.J.Siers and Md Zahidul Islam,"Software defect prediction using a cost sensitive decision forest and voting and a potential solution to the class imbalance problem," in Information Systems, vol. 51, 2015, pp. 62-71.
5. R.Xu, T.Chen, Y.Xia, Q.Lu,B. Liu, and X.Wang," Word embedding composition for data imbalances in sentiment and emotion classification," Congnitive Computation,vol.7, issue 2, 2015 ,pp. 226-240.
6. E.Ramentol, I.Gondres, S. Lajes, R.Bello,Y.Caballero, C.Cornelis and F.Herrera," Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance the SMOTEFRST -2T algorithm," Engineering Applications of AI, vol. 48, 2016, pp. 134-139.
7. B.Krawczyk, M.Galar,L. Jelen and F.Herrera," Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy," Applied Soft Computing, vol. 38,2016, pp.714-726.
8. A.Azaria, A. Richardson,S. Kraus and V.S. Subramanian," Behavioral analysis of insider threat: a survey and bootstrapped prediction in imbalanced data," IEEE Trans. Comput. Soc. Syst., vol.1, issue 2, 2014, pp. 135-155.
9. X.Goa, Z.Chen, S.Tang, Y. Zhang and J.Li," Adaptive weighted imbalance learning with application to abnormal activity recognition," Neurocomputing, vol.173, 2016, pp. 1927-1935.
10. M. Kubat, R.C. Holte, and S. Matwin, "Machine Learning for the Detection of Oil Spills in Satellite Radar Images," Machine Learning, vol. 30,issue 2-3,1998, pp. 195-215.
11. T. Fawcett and F. Provost, "Adaptive Fraud Detection," Data Mining and Knowledge Discovery, "vol. 3, issue 1, 1997, pp. 291-316.
12. N.Japkowicz and S.Stephen," The class imbalance problem: a systematic study," Intelligent Data Analysis, vol. 6, issue 5, 2002, pp.429-449.
13. Lopez, A. Fernandez, S. Garcia, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," Inf.Sci., vol. 250, 2013,pp. 113-141.
14. N. V. Chawia, K. W. Bowyer, L. O. Hall and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, 2002, pp. 321-357.
15. H. Han, W.-Y. Wang and B.-H. Mao, "Borderline-smote: A new oversapling method in imbalanced data sets learning," International Conference on Adv. Intelligent Computing, 2005, pp. 878-887.
16. J. Prusa, T. M. Khoshgoftaar, D. J. Dittman, and A. Napolitano, "Using random undersampling to alleviate class imbalance on tweet sentiment data," IEEE International Conference on Information Reuse and Integration, 2015, pp. 197-202.
17. S.-J. Yen and Y.-S. Lee, "Cluster-based under-sampling approaches for imbalanced data distributions," Expert Syst. Appl., vol.36, issue 3, 2009, pp. 5718-5727.
18. W.W. Ng, J. Hu, D. S. Yeung, S. Yin, and F. Roli, "Diversified sensitivitybased undersampling for imbalance classification problems," IEEE Trans.Cybern., vol.45, issue 11, 2015, pp.2402-2412.
19. N.Ofek, L. Rokach, R. Stern, and A. Shabtai, "Fast-CBUS: A fast clustering-based undersampling method for addressing the class imbalance problem," Neurocomputing, vol. 243, 2017, pp. 88-102.
20. T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," ACM SIGKDD Explorations Newsletter, vol. 6, issue 1, 2004, pp. 40-49.
21. S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE-majority weighted minority oversampling technique for imbalanced data set learning," IEEE Trans. Knowl. Data Eng., vol.26, issue 2, 2014, pp. 405-425.
22. M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting -, and hybrid-based approaches," IEEE Transactions Systems Man and Cybernetics-part c: Applications, vol.42, issue 4, 2012, pp.463-484.
23. N. Nikolaou, N. Edakunni, M. Kull, P. Flach, and G. Brown, "Cost sensitive boosting algorithms: Do we really need them?" Mach. Learn., vol. 104,issue 2-3,2016, pp. 359-384.
24. W. Benjamin and J. Nathalie, "Boosting Support Vector Machines for Imbalanced Data Sets," Foundations of Intelligent Systems, vol.4994, 2008, pp. 38-47.

AUTHORS PROFILE

Shaheen Layaq received her BCA (Bachelor Of Computer Application) in 2001, M.Sc(Information System) in 2003, B.Ed in 2012 from Kakatiya University, Warangal, Telangana, India. She has also completed her M.Tech (CSE) in 2011from JNTUH, Telangana, India and presently she is pursuing her Ph.D (Computer Science) from Kakatiya University, Warangal, Telangana, India under the guidance of Dr. B. Manjula. She also qualified her APSET in 2014. She is working as contract lecturer in Department of Computer Science, Singareni Collieries Women's Degree and PG College, kothagudem, Bhadradri kothagudem, Telengana, India. Her interested areas are Data Mining and Big Data Analytics. She has 16 years of teaching experience and published one research paper.

Dr. B. Manjula received her BCA (Bachelor Of Computer Application) and M.Sc(Information System) from Osmania University, Hyderabad, Telangana , India and Ph.D (Computer Science) from kakatiya University, Warangal. She is working as Assistant professor in Department of Computer Science, Kakatiya University, Warangal, TS, India. She is guiding the 8 research scholars. She is member of ACM and other professional organizations. She has published a book by title "Data Mining Techniques in Tracing the Trends of Business" in the year 2016. She has published more than 12 research papers and presented 10 research papers in international and national conferences in area of Data Mining, Big Data, Computer Networks and Neural Networks.