Experimental Proof of Unbiased Performance of B-mean Metric for Unbalanced Data

Pratik A Barot

(Email: Pratikabarot@gmail.com)

Abstract-*Unbalanced* data studies are still an open issue. Performance analysis and comparison of classifier is more challenging task in case of unbalanced data. Use of wrong and biased evaluation metric gives false result comparison. False comparison leads to false research finding. For unbalanced data classification unbiased performance evaluation is challenging task. We perform detail experiment to verify unbiased performance of newly proposed balanced mean parameter called B-mean for real-life dataset. Our result proved B-mean is more balanced metric as compared to existing most balanced metric called G-mean. We used eight real-life datasets and one synthetic dataset and proved that B-mean shows balanced performance evaluation even if imbalanced ratio is too high.

Keywords**:** *Unbalanced data classification, machine learning, evaluation metrics, B-mean*

I INTRODUCTION

Classifier performance is evaluated using evaluation metrics. Different researchers use different evaluation matrices. Selection of ideal evaluation metric for valid performance evaluation is always challenging task [2]. Selection of evaluation metrics for classifiers depends upon the data characteristics as well [2]. For balanced dataset most of the researchers prefer accuracy as performance evaluation metric. Balanced data have equal class distribution [3]. In case of equal class distribution accuracy gives overall performance result of classifier. However in case unbalanced data accuracy is not a good measure for performance valuation [4].

Unbalanced data does not have symmetric class distribution [8]. In unbalanced dataset some classes are in majority number and known as majority class and some are in minority number known as minority class [3]. Most of existing studies performed on unbalanced data gave more importance to majority class [3, 4]. They select accuracy, precision and recall metrics for performance evaluation. However, majority class is not important in unbalanced datasets like medical dataset, chemical reaction analysis, security system, weather forecasting, sentiment analysis, accident analysis etc. [1, 5, 6, 7].

Minority sensitive dataset need minority sensitive evaluation metric. In this paper we provide experimental study of performance analysis of new balanced mean called B-mean for real life datasets. B-mean is imbalanced ratio (IR) based metric which provide balanced performance evaluation. B-mean alleviate majority class biasing which is present in other evaluation matrices.

Table-1 shows evaluation matrices used in data mining [9]. In additional to matrices shown in Table-1 there is one more metric called area under curve (AUC). AUC is computed using graphic plot for true positive (TP) versus false positive (FP) [6].

TABLE-1 DATA MINING EVALUATION MATRICES [9]

Metrics Name	**Expression**
Accuracy	(TP + TN) / (TP+TN+FP+FN)
Precision	TP / TP+FP
Recall (TP rate or Sensitivity)	TP / TP + FN
TN rate (Specificity)	TN / TN + FP
Fvalue	$\frac{(1+\beta^2)Recall * Precision}{\beta^2 Recall + Precision}$ OR 2*((precision*recall)/(precision + recall))
Geometric Mean (G-mean)	$\sqrt[n]{(\prod_{i=1}^{n} ACi)}$

TABLE-2 DATASET DESCRIPTION

Name	**No of Attri.**	**No of Classes**	**No of Inst.**	**Remarks**
New-Thyroid	5	3	215	Hyper and Hypo are minority class. Normal is largest class with 150 instances.
Breast Cancer	10	2	699	Malignant is minority class
HIV	9	2	1625	Imbalanced dataset
Contraceptive	10	3	1473	Survey dataset
Contact Lenses	4	3	24	One class in in majority and two are in minority
Diabetes	9	2	768	Imbalanced dataset
Vote	17	2	435	Imbalanced dataset
Synthetic	10	2	1000	Created synthetically

II METHOD AND RESULT ANALYSIS

We used eight datasets from UCI repository and one synthetic dataset. Dataset description is shown in Table-2.

We used Weka library for our experiment. We applied naïve Bayesian classifier on eight datasets and analyze the classifier performance using matrices of Table-1 and compare their result with B-Mean value. B-mean value is calculated using (1).

B-mean =

$$(((IR\times TN\_rate)+(1/IR\times TP\_rate) \div (IR+1/IR))+Acc)\div 2 \quad (1)$$

Table-3 shows classifier result in form of different evaluation measures. In this table P is for positive instances and N is for Negative instances. In case of multi-class dataset we considered all majority classes as positive and all minority classes as negative.

From Table-3 result we discovered that G-mean is more balanced as compared to Precision and Recall. Although G-mean is termed as balanced metric, we found that it has little biasing towards majority class. For Vote dataset majority class have less accuracy (TP_rate) as compared to minority class accuracy (TN_rate) and thus G-mean is less than B-mean. For all other datasets where majority class accuracy is more than the minority class, the G-mean value is more than the B-mean value.

Fig.1 shows behavioral graph of evaluation matrices. If imbalanced ratio is comparatively less and TP_rate and TN_rate do not much differ then there is no significant difference between G-mean and B-mean value. However, if there is major difference between TP_rate and TN_rate then the B-mean is more balanced as compared to the G-mean. As imbalanced ratio increases G-mean have visible biasing and B-mean is proved as more balanced metric.

TABLE-3 CLASSIFICATION RESULTS

P	N	TP	TN	FP	FN	Acc	TN_rate	TP_rate	B-mean	Gmean	Dataset
458	241	443	237	4	15	0.97	0.98	0.97	0.98	0.98	BC
1250	375	1201	322	53	49	0.94	0.86	0.96	0.90	0.91	HIV
150	65	150	64	1	0	1.00	0.98	1.00	0.99	0.99	NT
1140	333	867	171	162	273	0.70	0.51	0.76	0.62	0.62	Contraceptive
15	9	12	5	4	3	0.71	0.56	0.80	0.66	0.67	Contact -Lenses
700	300	605	149	151	95	0.75	0.50	0.86	0.65	0.66	Credit
500	268	422	164	104	78	0.76	0.61	0.84	0.71	0.72	Diabetes
267	168	238	154	14	29	0.90	0.92	0.89	0.91	0.90	Vote
800	200	700	43	167	100	0.74	0.20	0.88	0.49	0.42	Synthetic

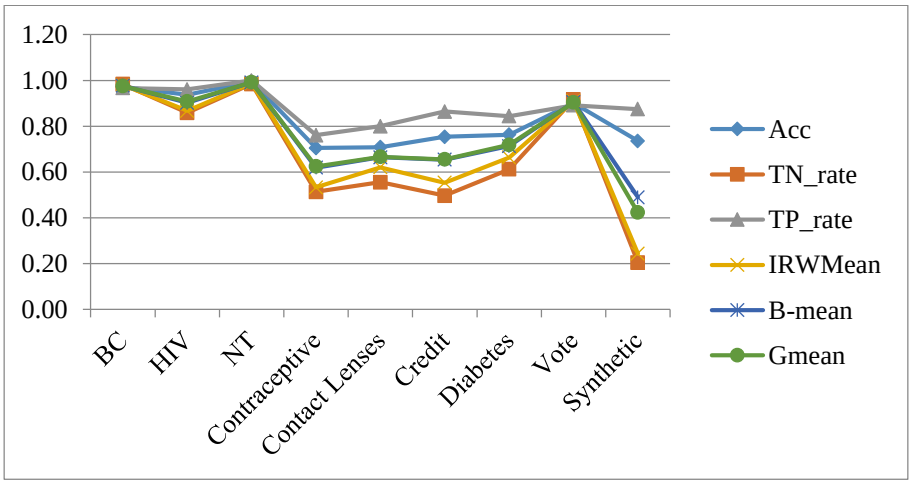

Fig.1 Result of Matrices for Unbalanced data

Fig.1 shows than in most of the time B-mean and G-mean are overlapping. However, in case of synthetic dataset where imbalanced ratio is high and there is major difference between TP_rate and TN_rate the B-mean gives balanced result as compared to the G-mean.

III CONCLUSION

Balanced evaluation is important for performance evaluation of classification. Most of existing studies uses Accuracy, Precision, Recall, F-measure, G-mean as evaluation metric. Except G-mean all other are majority class biased matrices. We evaluate performance of B-mean for real-life datasets and found that B-mean is more balanced as compared to G-mean. As imbalanced ratio increases the difference between G-mean and B-mean increases and B-mean shows clear upper-hand over G-mean.

REFERENCES

[1] Bartosz Krawczyk1 (2016) Learning from imbalanced data: open challenges and future directions, Prog Artif Intell, Springer

[2] Jerzy Stefanowski (2016) Dealing with Data Difficulty Factors While Learning from Imbalanced Data, Springer International Publishing Switzerland.

[3] Pratik A. Barot, H. B. Jethva (2017) Statistical Study to Prove Importance of Causal Relationship Extraction in Rare Class Classification, (ICTIS 2017) - Volume 1, Smart Innovation, Systems and Technologies Springer DOI 10.1p0ra0t7ik/9a7b8a-r3o-t3@19g-m63a6il7.c3o-3m_51

[4] Astha Agrawal, Herna L Viktor, Eric Paquet (2016) SCUT: Multi-Class Imbalanced Data Classification using SMOTE and Cluster-based Undersampling, In: Proceedings of the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2015) - Volume 1: KDIR, pages 226-234, SCITEPRESS – IEEE explore.

[5] Randa Oqab Mujalli, Griselda Lopez, Laura Garach, Bayes Classifiers for Imbalanced Traffic Accidents Datasets, Accident Analysis and Prevention, Elsevier, Dec-2015.

[6] Pratik A. Barot, H. B. Jethva (2018) Enhance Decision Tree algorithm for Unbalanced Data: RareDTree. International Journal of Computer Engineering and Technology, pp. 109-115. http://www.iaeme.com/IJCET/issues.asp?JType=IJCET&VType=9&IType=5

[7]Dongmei Zhang, Jun Ma, Jing Yi, Xiaofei Niu, Xiaojing Xu, An Ensemble Method for Unbalanced Sentiment Classification, in: proceeding of 11th International Conference on Natural Computation, IEEE-2015.

[8]N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: Synthetic minority over-sampling technique, Journal of Artificial Intelligence Research 16 (2002) 321–357.

[9] Jiawei Han, M Kamber, J Pei, Data Mining Concepts and Techniques; Third Edition,Elsevier 2012.