A Novel Technique on Class Imbalance Big Data using Analogous under Sampling Approach

Mohammad Imran
Regd No: PP.COMP.SCI&ENG.0308C, Research Scholar, Computer Science and Engineering, Rayalaseema University, Kurnool-518007, Andhra Pradesh, India

Vaddi Srinivasa Rao
Ph.D, Professor & Head Department of Computer Science and Engineering, Department of CSE, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada – 520007,

ABSTRACT

In this paper, we propose hybrid Random under Sampled Imbalance Big Data (USIBD) framework to extract knowledge from class imbalance big data. A novel under-sampling method for the base learner is also proposed to handle the dynamic class-imbalance problem caused by the gradual evolution of classes in big data. The proposed USIBD knowledge discovery framework is robust and less sensitive to outliers where non-uniform distribution of data is applied. Empirical studies demonstrate the effectiveness of USIBD in various class imbalance big datasets scenarios in comparison to existing methods.

Keywords

Classification, Big data, Imbalanced data, Under Sampling, USIBD

1. INTRODUCTION

Data mining is the process of discovering hidden knowledge from the existing databases. The main approaches are classification, clustering, association analysis and pattern mining etc.

Classification is the process of classifying the instances in the existing labeled classes by analyzing the features of the instances [1].

The most popular classification techniques are decision trees, neural networks, support vector machines etc. In clustering, the instances are formed as clusters or groups depending upon the intrinsic properties of the instances. The popular clustering approaches are k-means, DB-scan, Hierarchical clustering etc.

In classification one of the effective and efficient approaches is decision trees. The decision trees are formed by the process of induction. The training instances are used to build the decision tree model and the testing subset is used to assess the performance of the build decision tree on the unseen instances. The class imbalance datasets are one of the new data source emerged recently. In binary class imbalance datasets, there exist two sub classes; majority and minority.

The majority subclass is the one in which large percentage of instances from one class exists. In minority subset only less percentage of instances from other class exists. The performance of the existing classification drastically degrades when applied to class imbalance datasets. The reason for reduced performance is due to improper model built with the training instances. Since in the training subset, only few minority instances are available for model building. The model is very weak to predict the unseen minority instances.

The class imbalance problem also shows its presence in the case of big data sources in real time. In the context of big dataset the reduced performance is seen in the classification algorithms for class imbalance data. A new series of novel approaches are needed to address the problem of class imbalance on big data.

2. RELATED WORK

Many algorithms and methods have been proposed to ameliorate the effect of class imbalance on the performance of learning algorithms.

Rajiv Sambasivan et al [2] have presented an algorithm for classification tasks on big data which is as accurate as ensemble methods such as random forests or gradient boosted trees. Unlike ensemble methods, the models produced by the algorithm can be easily interpreted. Petra Perner [3] have developed a method that allows automatically to discover the decision rules for diagnosing medical images in normal tissue images and images showing a polyp. Tianyi Yang et al [4] have implemented HDFS and Map Reduce for a well-known learning algorithm—decision tree in a scalable fashion to large input problem size.

Armando Segatori et al [5] have propose a distributed FDT learning scheme shaped according to the Map Reduce programming model for generating both binary and multi-way FDTs from big data. The scheme relies on a novel distributed fuzzy discretizer that generates a strong fuzzy partition for each continuous attribute based on fuzzy information entropy.

Hanif Arief Wisesa et al [6] have presented a comparison of processing large traffic data by using decision trees implemented in MoA and SPARK MLib, which successfully regress the traffic dataset as a data stream quickly with a fairly good accuracy.

3. FRAMEWORK OF USIBD ALGORITHM

The algorithm Under Sampled Imbalance Big Data (USIBD) learning is a unique framework, which performs under sampling by following a strategic approach of removing the instances from the majority subset. Under sampling can help improve run time and storage problems by reducing the number of training data samples when the training data set is huge.

These limitations are uniquely addressed in our proposal such as: under sampling can discard potentially useful information which could be important for building rule classifiers.

The sample chosen by random under sampling may be a biased sample. It will not be an accurate representative of the population and thereby, resulting in inaccurate results with the actual test data set.

In different scenarios, an aim of under sampling is to balance class distributions. The process of eliminating majority instances depending upon unique properties of the datasets can be extended for different percentages.

Our proposed method consists of two steps. In the first step, we construct an influence space around a test point p. In the second step a rank difference based outlier score is assigned on the basis of this influence space.

3.1 Influence Space Construction

Influence space depicts a region with significantly high reverse density in the locality of a point under consideration.

If the localities of the neighbours within the influence space are denser with respect to the locality of the concerned point, then a high value of outlierness core will be assigned to it. For an entire dataset, number of neighbours in the influence space is kept fixed. As the distance is increased from the target point, more number of neighbours gets included in its surroundings result

In given different values of radius R, with successive addition of neighbouring points, a set of reverse densities is obtained for each point at varying depths (number of neighbouring points). The average reverse density R for each depth is determined next. Note that we have considered the depth and not the distance around the neighbours to handle situations where there is empty space (no neighbouring point is present) surrounding a given point. To avoid random fluctuations, the variation in the average reverse density with respect to depth has been smoothed using a Gaussian kernel.

In this smoothing process, an optimal width for the kernel optimal is determined using better estimation of the significant density fluctuation around the neighbor points. We deem the first most significant peak in this smoothed kernel probability density function as the limit of the influence space. The peak has been determined using the undecimated value.

3.2 Outlier score

In the second part of our proposed algorithm we have used a rank difference based score for ranking of the outliers. The positive

value of the rank difference (R−k) signifies the high concentration of the neighbours around the training point q than that of the test point p. The negative and zero value respectively signify a lower or same concentration of the training points around q than that of p. Thus the outlierness of the test point depends directly on the excess population of the neighbourhood space of q with respect to the test point p, i.e., on the rank difference (R−k). Secondly, it also depends inversely on its own forward density.

4. DATASETS

Experiments are conducted using eight datasets from UCI [7] data repositories. Table 1 summarizes the benchmark datasets used in the anticipated study. For each data set, S.no., Dataset, name of the dataset, Instances, number of instances, Attributes, Number of Attributes, IR, Imbalance Ratio are described in the table for all the datasets.

Table 1 UCI datasets and their properties

S.no.	**Dataset**	**Inst**	**Attributes**	**IR**
1.	Car	1728	7	18.61
2.	German_credit	1000	21	2.33
3.	Hypothyroid	3772	30	36.64
4.	Mfeat	2000	217	9.0
5.	Nursery	12960	9	13.17
6.	Page-blocks	5473	11	14.93
7.	Segment	2310	20	6.0
8.	Sick	3772	30	15.32

We performed the implementation of our new algorithms within the Weka [8] environment on windows 7 with i5-2410M CPU running on 2.30 GHz unit with 4.0 GB of RAM. The validation of the results is done using 10 fold cross validation, in which the dataset is split into 10 subsets and in each run nine subset are used for training and the remaining subset is used for testing. In 10 runs, the testing subset is altered and average measures for the 10 runs are generated. The evaluation metrics used in the paper are detailed below,

Accuracy is the percentage of correctly classified instances. AUC can be computed simple as the micro average of TP rate and TN rate when only single run is available from the clustering algorithm.

The Area under Curve (AUC) measure is computed by,

$$\boxed{AUC = \frac{1+TP_{RATE}-FP_{RATE}}{2}}$$

-----------------(1)

[Or]

$$\boxed{AUC = \frac{TP_{RATE}+TN_{RATE}}{2}}$$ ------------------ (2)

The Precision measure is computed by,

$$\boxed{\Pr ecision = \frac{TP}{(TP)+(FP)}}$$ ------------------ (3)

The Recall measure is computed by,

$$\boxed{\mathrm{Re}\,call = \frac{TP}{(TP)+(FN)}}$$ -----------------(4)

5. EXPERIMENTAL RESULTS

In the experimental setup, we have considered 8 datasets from UCI repository, which are in large size (few thousand instances). The validation is done by using the 10 fold cross validation for 10 runs.

The mean of all the measures for 10 runs is used as experimental results. We have compared our proposed USIBD algorithm with C4.5 [9] algorithm which is one of the benchmark algorithms in decision trees. The reported experimental results suggest that our proposed algorithm has performed well than the existing C4.5 algorithm.

The validation measures used in the experimental simulation are AUC, Precision, Recall and F-measure. In validation measures AUC, Precision Recall and F-measure there is an increase in the values are reported for an improved performance.

If the proposed USIBD algorithm is better than the compared technique then '●' symbol appears in the column.

If the proposed USIBD algorithm is not better than the compared technique then '○' symbol appears in the column.

Table 2 reports the results of our proposed USIBD algorithm verse C4.5 algorithm in terms of AUC. The AUC values generated by USIBD algorithm are improved than C4.5 algorithm on 3 out of 8 datasets.

Table 3 reports the results of our proposed USIBD algorithm verse C4.5 algorithm in terms of precision.

The precision values generated by USIBD algorithm are improved than C4.5 algorithm on 5 out of 8 datasets.

Table 4 reports the results of our proposed USIBD algorithm verse C4.5 algorithm in terms of recall.

The recall values generated by USIBD algorithm are improved than C4.5 algorithm on 2 out of 8 datasets.

The mean performances were significantly different according to the T-test at the 95% confidence level.

Table 2 Summary of tenfold cross validation performance for AUC on all the datasets

Datasets	C4.5	USIBD
anneal	0.931±0.164●	0.938±0.166
car	0.981±0.011○	0.919±0.080
cmc	0.691±0.049●	0.692±0.048
kr-vs-kp	0.998±0.003○	0.998±0.002
letter	0.985±0.011○	0.983±0.012
mfeat	0.967±0.036●	0.969±0.030
mushroom	1.000±0.000	1.000±0.000
nursery	1.000±0.000	1.000±0.000

Table 3 Summary of tenfold cross validation performance for Precision on all the datasets

Datasets	C4.5	USIBD
anneal	0.505±0.500●	0.660±0.454
car	0.972±0.016○	0.923±0.131
cmc	0.606±0.051●	0.613±0.048
kr-vs-kp	0.994±0.006●	0.995±0.006
letter	0.952±0.028●	0.953±0.022
mfeat	0.921±0.077●	0.935±0.065
mushroom	1.000±0.000	1.000±0.000
nursery	1.000±0.000○	0.400±0.492

Table 4 Summary of tenfold cross validation performance for Recall on all the datasets

Datasets	C4.5	USIBD
anneal	0.510±0.502●	0.700±0.461
car	0.962±0.018○	0.771±0.176
cmc	0.617±0.063○	0.614±0.068
kr-vs-kp	0.995±0.005○	0.994±0.007
letter	0.965±0.023○	0.961±0.024
mfeat	0.925±0.080●	0.938±0.062
mushroom	1.000±0.000	1.000±0.000
nursery	1.000±0.000○	0.400±0.492

Table 5 Summary of tenfold cross validation performance for F-measure on all the datasets

Dataset	C4.5	USIBD
anneal	0.507±0.500 ●	0.673±0.452
car	0.967±0.011○	0.827±0.135
cmc	0.610±0.049 ●	0.612±0.048
kr-vs-kp	0.995±0.004○	0.994±0.004
letter	0.958±0.021○	0.957±0.017
mfeat	0.921±0.069 ●	0.935±0.053
mushroom	1.000±0.000	1.000±0.000
nursery	1.000±0.000○	0.400±0.492

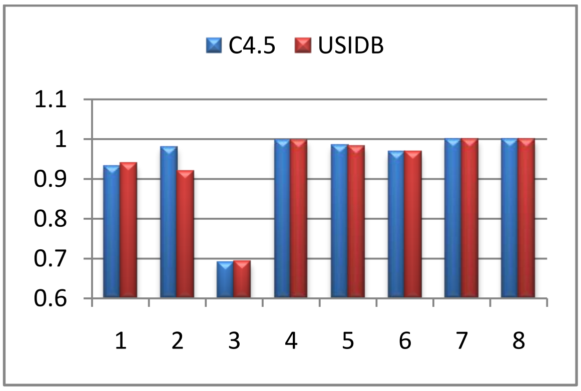

Fig. 1 Trends of USIBD v/s C4.5. on imbalance Big dataset

Figure.1 presents the summary of the experimental results of USIBD algorithm verse C4.5 algorithm on different evaluation metrics. The registered wins of USIBD algorithm on C4.5 shows that our proposed algorithm is better than the existing algorithm on class imbalance datasets.

Finally, we can say that USIBD is one of the best alternatives to handle class imbalance problems effectively.

This experimental study supports the conclusion that a prominent recursive over sampling approach can improve the CIL behavior when dealing with imbalanced datasets, as it has helped the USIBD methods to be the best performing algorithms when compared with C4.5 algorithm.

6. CONCLUSION

As new data and updates are constantly arriving, the results of data mining applications become stale and obsolete over time. Incremental under sampling is a promising approach to refreshing mining results. It utilizes previously saved states to avoid the expense of re-computation from scratch. In this paper, we propose hybrid Radom Under Sampled Imbalance Big Data (USIBD) to extract knowledge from class imbalance big data. A novel under-sampling method for the base learners is also proposed to handle the dynamic class-imbalance problem caused by the gradual evolution of classes in big data. The proposed USIBD knowledge discovery framework is robust and less sensitive to outliers where non-uniform distribution of data is applied. Empirical studies demonstrate the effectiveness of USIBD in various class imbalance big datasets scenarios in comparison to existing methods.

7. REFERENCES

[1] O. Maimon, and L. Rokach, *Data mining and knowledge discovery handbook*, Berlin: Springer, 2010.

[2] Rajiv Sambasivan, SourishDas,"Big Data Classification Using Augmented Decision Trees", arXiv preprint arXiv:1710.09567, 2017.

[3] Petra Perner,"Big Data, Decision Tree Induction, and Image Analysis for the Discovery of Decision Rules for Colon Examination", International Journal of Engineering Research & Science (IJOER) ISSN: [2395-6992] [Vol-3, Issue-8, August- 2017].

[4] Tianyi Yang and Anne HeeHiongNgu,"Implementation of Decision Tree Using Hadoop Map Reduce",Yang and Ngu, Int J Biomed Data Min 2016, 6:1

a. DOI: 10.4172/2090-4924.1000125.

[5] Armando Segatori, Francesco Marcelloni, and Witold Pedrycz," On Distributed Fuzzy Decision Trees for BigData",DOI10.1109/TFUZZ.2016.2646746,IEEE Transactions on Fuzzy Systems.

[6] Hanif Arief Wisesa, M. Anwar Ma'sum, PetrusMursanto, Andreas Febrian,Processing Big Data with Decision TreesA Case Study in Large Traffic Data", IWBIS 2016 978-1-5090-3477-2/16/2016 IEEE.

[7] Blake C, Merz CJ (2000) UCI repository of machine learning databases. Machine-readable data repository. Department of Information and Computer Science, University of California at Irvine, Irvine.http://www.ics.uci.edu/mlearn/MLRepository.html .

[8] Witten, I.H. and Frank, E. (2005) Data Mining:Practical machine learning tools and techniques.2nd edition Morgan Kaufmann, San Francisco.

[9] J. Quinlan. C4.5 Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann, 1993.

8. AUTHOR'S PROFILE

Mohammad Imran received his B.Tech (CSE) and M.Tech (CSE) in 2008 from JNTU, Hyderabad, His Research interests include Big Data Analytics, Artificial Intelligence, Class Imbalance Learning, Ensemble learning, Machine Learning and Data mining.
He is a Research Scholar with **Regd No: PP.COMP.SCI&ENG.0308C** in the department of Computer Science and Engineering, **Rayalaseema University**, **Kurnool-518007, Andhra Pradesh**. He is currently working as an Assistant Professor in Department of CSE, Muffakham Jah College of Engineering and Technology, Banjara Hills, Hyderabad-500034, India. You can reach him at imran.quba@gmail.com.

Dr.Vaddi Srinivasa Rao, Professor & Head Department of Computer Science and Engineering, Velagapudi Ramakrishna Siddhartha Engineering College, Vijayawada-520007,Andhra Pradesh, India, His Research Interest Includes Big Data Analytics, Computer Networks, Information Security, Artificial Intelligence, Class Imbalance Learning, Ensemble learning, Machine Learning and Data mining.