Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 11, November 2015, pg.338 – 343

SURVEY ARTICLE

A Survey on Methods to Handle Imbalance Dataset

Apurva Sonak[1], R.A.Patankar[2]

[1]Computer Science, MIT, India

[2]Computer Science, MIT, India

[1] apurvasonak26@gmail.com; [2] ruhi.patankar@mitpune.edu.in

Abstract - Imbalanced data set, a problem often found in real world application, can cause seriously negative effect on classification performance of machine learning algorithms. There have been many attempts at dealing with classification of unbalanced data sets. To handle the problem of imbalanced data is to re balance them artificially by oversampling and/or under-sampling.

Keywords – Imbalance dataset, classifiers, sampling, cost-sensitive learning, Data mining

I. INTRODUCTION

Imbalanced data sets are a special case for classification problem where the distribution of class is not equal among the classes. Typically, there are two classes: The majority i.e. negatives class and the minority i.e. positive class. These type of data suppose a new challenging problem for Data Mining, since standard classification algorithms usually consider a balanced training set and this supposes a bias towards the majority (negative) class. The imbalanced data set problem appears in many real world applications like text categorization, fault detection, fraud detection, oil-spills detection in satellite images, toxicology, cultural modeling, and medical diagnosis.

II. IMBALANCE PROBLEM

The first step in providing viable solutions for imbalanced domains is to understand the problem: what is the real issue with the imbalance? Initially, the difficulty of dealing with imbalance problems was thought of

coming from its imbalance rate (IR), i.e. the ratio between the number of instances in the majority (mMaj) and minority classes (mMin).The problem that occurs when the data is imbalanced are:

•The cost of missing a minority class is typically much higher that missing a majority class.

•Most learning systems are not prepared to cope with large difference between the numbers of cases belonging to each class.

•Classification algorithm underperforms when data is unbalanced.

The imbalance problem is a relative problem, which depends on: (1) The imbalance ratio, i.e. the ratio of the majority to the minority instances, (2) The complexity of the concept represented by the data, (3) The overall size of the training set and (4) The classifier involved. The issue with the imbalanced learning problem is the ability of imbalanced data to significantly compromise the performance of most standard machine learning algorithms. Most algorithms usually assume balanced class distributions. This problem will cause most standard machine learning algorithms to be biased toward the majority class because they try to optimize overall accuracy, which is overwhelmed by majority classes and ignore minority class. The methods proposed for dealing with data imbalance include both data and algorithmic levels. Data level solutions include resampling the data. In next section will we discuss these methods[1].

III. SAMPLING METHODS

Easy method for balancing the imbalanced data set is sampling the data. Sampling involves oversampling and under sampling.

A. Oversampling

Oversampling is a sampling technique which balances the data set by replicating the examples of minority class. It is also called up sampling. The advantage of this method is that there is no loss of data. The disadvantage of this technique is it may lead to over fitting and can introduce an additional computational overhead. Oversampling is also divided into two types.:Random Oversampling and Informative Oversampling. Random Oversampling is the method which balances the class distribution by replicating the randomly chosen minority class examples. Informative Oversampling method synthetically generates minority class examples based on a pre-specified criterion[2].

Random oversampling is a simple yet effective approach to resampling. In this, one chooses members from the minority class at random; these randomly chosen members are then duplicated and added to the new training set. One must note two things when randomly oversampling: First, one chooses documents randomly from the original training set, not the new training set. Second, one always randomly oversamples with the replacement. If we were to randomly oversample without replacement, we would deplete all members of the minority class long before we reached the desired balance between the majority and minority class**.**

There are number of Oversampling methods available in the literature like SMOTE, Borderline SMOTE, OSSLDDD-SMOTE etc. SMOTE is an oversampling method which create "synthetic" example rather than oversampling by replacements. The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any/all of the k minority class nearest neighbors. Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen.

B. Undersampling

Under-sampling is an efficient method for balancing data. This method uses a subset of the majority class to train the classifier. In undersampling, we delete some examples of the majority class. Undersampling methods are divided into random and informative. Random Undersampling is simple, it randomly eliminates samples from the majority class till the data set gets balanced. Informative Undersampling method selects only the required majority class examples based on a pre-specified selection criterion to make the data set balanced.

Informative Undersampling can be passive or active. Passive selection methods are proposed as preprocessing technique for selecting informative samples for a classifier. In Active selection methods, informative samples are queried during the construction process of the classifier .The most common preprocessing technique is random majority under-sampling (RUS). IN RUS, Instances of the majority class are randomly discarded from the dataset. However, the main drawback of under-sampling is that potentially useful information contained in these ignored examples is neglected. There many ways attempts to improve upon the performance of random sampling, such as Tomek links, Condensed Nearest Neighbor Rule and One-sided selection etc[5].

Tomek link can be defined as follows: Consider the two examples a and b which belongs to different classes, and d (a,b) is the distance between a and b. A(a,b) pair is called a Tomek Link if there is not an example c, such that $d(a,c)<d(a,b)$ or $d(b,c)<d(a,b)$. If two examples form a Tomek link, then either one of these examples is noise or both examples are border-line and then we can discard the examples.

IV. COST SENSITIVE LEARNING

At the algorithmic level, solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the probabilistic estimate at the tree leaf (when working with decision trees), adjusting the decision threshold, and recognition-based (i.e., learning from one class) rather than discrimination-based (two class) learning. Cost Sensitive Learning (CSL) is another commonly used approach to handle the classification problem of imbalanced data sets. Cost sensitive learning apply the miss classification cost to incorrectly classified examples. For correct classification there is no penalty. The cost of FN will be more than the cost of FP and the costs of TP and TN is zero.

		Predicted	
		Positive	Negative
Actual	Positive	0	C(FN)
	Negative	C(FP)	0

There are many ways to implement cost sensitive learning, it is categorized into three, the first class of techniques apply misclassification costs to the data set as a form of data space weighting, the second class applies cost-minimizing techniques to the combination schemes of ensemble methods, and the last class of techniques incorporates cost sensitive features directly into classification paradigms to essentially fit the cost sensitive framework into these classifiers.

The two classes data distribution is unequal which says that it is imbalanced data set. The type of learning algorithm which takes misclassification cost into consideration is called Cost Sensitive Learning(CSL). It produces the classifier with minimum total cost. The advantage of this method is here no data is replicated or eliminated.

V. COMPARISON

Sampling and cost sensitive are the two methods to handle imbalanced data set. Performance of these methods depends on what dataset we are using. The factors affecting on the performance of these methods are 1) Size of the data set, 2) imbalance ratio of the classes in the dataset. The table shown below show which method will perform best and worst in the given cases:

	Oversampling	**Undersampling**	**Cost- Sensitive**
Size of dataset is Large	Worst	Worst	Best
Size of dataset is Small	Best	Better	Worst
Disadvantages	Increases the Processing time.	Loss of data	Need to give the cost for missclassification.

VI. LITERATURE SURVEY

Pengyi Yang, Paul D. Yoo, Juanita Fernando, Bing B. Zhou, Zili Zhang, and Albert Y. Zomaya proposed a new SSO(sample subset optimization technique in Sample Subset Optimization Techniques for Imbalanced and Ensemble Learning Problems in Bioinformatics Applications[1].

Dr.D.Ramyachitra, P.Manikandan discussed the problems of imbalanced dataset and there solutions in Imbalance dataset classification and solution[2]. M.Akhil jabbar , Dr.Priti Chandra, Dr.B.L Deekshatulu have done the experiments on Heart disease using kNN and genetic algorithm[6].

Atul Kumar Pandey, Prabhat Pandey, K.L. Jaiswal, Ashish Kumar Sen discussed the different attribute selection methods for data in Data Mining Clustering Techniques in the Prediction of Heart Disease using Attribute Selection Method. Vaishali Ganganwar gives an overview of the classification algorithm that can be used for imbalanced dataset in paper An Overview of Classification algorithms for Imbalanced dataset [5].

Feature Selection in Imbalance data sets, Ilnaz Jamali , Mohammad Bazmara and Shahram Jafari paper gives the techniques of feature selection[7].

Haibo He, Edwardo A. Garcia in paper" Learning from Imbalanced Data", explains the concept of imbalance data and the problems with them[14].

Sr No	**Paper**	**Year**	**Contents**
1	Imbalance dataset classification and Solutions: A Review	International Journal of Computing and Business Research,2014	Imbalance classification techniques, methods and the algorithms used to solve the problem of inequality .
2	An Overview of Classification algorithms for Imbalanced dataset	International Journal of Emerging Technology and Advanced Engineering ,2012	Data level solutions,Algorithmic level solution,Cost Sensitive learning.
3	On the Classification of Imbalanced Datasets.	International Journal of Computer Science & Technology,2011	Data level sampling and Cost Sensitive strategies
4	Sampling Approaches for Unbalanced Data Classification Problem.	2011	Over sampling and under sampling methods.
5	The performance of Resampling Strategies for the Class Imbalance Problem	2010	Oversampling performs best on the small dataset and worst on large dataset .

6	A Hybrid Approach to Learn with Imbalanced Classes using Evolutionary Algorithms.	International Conference on Computational and Mathematical Methods in Science and Engineering,2009	In this paper they Created a balance data using several minority classes and Combined the different classifiers.
7	Learning from Imbalance Data.	IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,2009	Cluster based sampling,Confusion matrix.
8	Cost-Sensitive Learning vs. Sampling: Which is Best for Handling Unbalanced Classes with Unequal Error Costs?"	2005	There is no clear winner but Cost sensitive performs better when cost is known .
9	SMOTE: Synthetic minority over-sampling technique	2002	Oversamples dataset by creating "synthetic" examples. Uses kNN algorithm

VII. CONCLUSION

Data imbalance is the most common problem. Existing classification algorithms underperform on the imbalance data, so we need to preprocess the data and make it balanced. Methods for making data balance are Sampling and Cost sensitive learning. At data level, sampling is the most common approach to deal with imbalanced data. over- sampling clearly appears as better than under-sampling for local classifiers, whereas some under-sampling strategies outperform over-sampling when employing classifiers with global learning.

VIII. FUTURE WORK

Future work is to build a new classifier which will perform better on class imbalance as the existing classifier underperform on this. We observed that current research in imbalance data problem is moving to hybrid algorithms.

REFERENCES

[1] Pengyi Yang, Paul D. Yoo, Juanita Fernando, Bing B. Zhou, Zili Zhang, and Albert Y. Zomaya,"Sample Subset Optimization Techniques for Imbalanced and Ensemble Learning Problems in Bioinformatics Applications", IEEE TRANSACTIONS ON CYBERNETICS, VOL. 44, NO. 3, MARCH 2014

[2] D.Ramyachitra, P.Manikandan, "Imbalanced Dataset Classification and Solutions: A Review", International Journal of Computing and Business Research (IJCBR) ISSN (Online) : 2229-6166 Volume 5 Issue 4 July 2014

[3] Beant Kaur,Williamjeet Singh," Review on Heart Disease Prediction System using Data Mining Techniques", International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 2 Issue: 10 **,**2014

[4] M.Akhil jabbar, B.L Deekshatulu ,Priti Chandra," Classification of Heart Disease Using K- Nearest Neighbor and Genetic Algorithm", ScienceDirect Procedia Technology 10 85 – 94,2013

[5] Vaishali Ganganwar," An overview of classification algorithms for imbalanced", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume 2, Issue 4, April 2012

[6] M.Akhil jabbar , Dr.Priti Chandra, Dr.B.L Deekshatulu," Heart Disease Prediction System using Associative Classification and Genetic Algorithm", International Conference on Emerging Trends in Electrical, Electronics and Communication Technologies-ICECIT, 2012

[7] Ilnaz Jamali , Mohammad Bazmara and Shahram Jafari," Feature Selection in Imbalance data sets", IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 2, May 2012

[8] Vidushi Sharma, Sachin Rai, Anurag Dev," A Comprehensive Study of Artificial Neural Networks", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, Issue 10, October 2012
[9] .V. KrishnaVeni,T. Sobha Rani," On the Classification of Imbalanced Datasets" IJCST Vol . 2, SP 1, December 2011
[10] . Maruthi Padmaja ,"Sampling Approaches for Unbalanced Data Classification Problem", thesis-2011.
[11] Hai Yun, M A Nan, Ruan Da, A N Bing," An Effective Oversampling Method For Imbalanced Data Sets Classification", 2011.
[12] Cente Garcia, Jose Salvador Sanchez, Ramon A. Mollineda, "Exploring the performance of Resampling Strategies for the Class Imbalance Problem",2010
[13] Nitesh V. Chawla, "Data Mining for Imbalanced Datasets: An Overview"
[14] Haibo He, Edwardo A. Garcia," Learning from Imbalanced Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 9, SEPTEMBER 2009
[15] Claudia Regina Milar_e1, Gustavo E.A.P.A. Batista1 and Andr_e C.P.L.F. Carvalho1," A Hybrid Approach to Learn with Imbalanced Classes using Evolutionary Algorithms", CMMSE 2009

