Novelty Journals

ISSN 2394-7314

International Journal of Novel Research in Computer Science and Software Engineering
Vol. 3, Issue 1, pp: (27-33), Month: January-April 2016, Available at: www.noveltyjournals.com

Analyzing the Effect of Preprocessing in Class Imbalanced Data

[1]N.Nandhini, [2]S.Poongothai, [3]M.Priya, [4]R.Priyadharsini, [5]Ms.N.Premalatha

[5]M.E. Assistant Professor, [1,2,3,4,5] Information Technology, Dr. Mahalingam College of Engineering and Technology
Pollachi, India

Abstract: **Class imbalance is one of the critical problems in classification. Because when classifiers classifies the records in imbalanced datasets, the classified datasets provides best classification accuracy for class(es) which has more number of examples and poor classification accuracy for the class(es). The poor classes have less number of examples. Therefore solving the class imbalance before applying the classification algorithm or at the time of applying the classification algorithm is very important, if good classification accuracy for minority class (es) (class with less number of examples) is needed. To identify the better approach to solve class imbalance problem, in this project the algorithm is compared with performance of four ensemble methods (Adaboost, bagging, random subspace & rotation forest). To evaluate the performance of the algorithm the datasets with severe, less severe imbalance ratios are used.**

Keywords: **Class imbalance, Majority class, minority class, Classification, Ensemble.**

1. INTRODUCTION

Class imbalance problem exist in data mining. In imbalanced datasets, the classes with more examples are majority classes and the one having fewer examples the minority classes. Class imbalance problem mostly occurs in classification problem. When the algorithm is applied to the dataset, the majority class instances are considered for classification and neglecting the minority class. It reduces the overall accuracy of the result which affects the future predictions. Hence this problem, existing in the datasets must be solved for obtaining high accuracy.

A simple example for class imbalance problem is Consider the two class dataset with ratio of two classes is 10:90. Two classification algorithms are introduced. The first classification algorithm gives overall predictions accuracy as 90% with accuracy of each class 0% and 100% respectively. The second classifications algorithm gives overall predictions accuracy as 78% with accuracy for each class is 60% and 80% respectively. Here, the first classification algorithm gives higher overall predictions accuracy than second classifications algorithm. But first classification algorithm misclassifies all the instances in class1. So in this cases, algorithms like first classification algorithm cannot be considered as a good classification algorithm.

1.1 Two class imbalance:

In two class imbalance, there are only two class (i.e.) Majority class and Minority class. When applying classification algorithms on different datasets, only majority class instances are correctly classified, whereas minority class instances are neglected.

1.2 Multi class imbalance:

Multi class imbalance is also similar to that of two class imbalance but in which majority classes easy to identified whereas minority classes difficult to identified. When classification algorithm is applied only majority class is identified, according to the algorithm applied minority class instances are not taken into consideration. So it is necessary for balancing the dataset affected by class imbalance problem.

2. LITERATURE SURVEY

2.1 LITERATURE SURVEY:

Weighted extreme learning machine for imbalance learning [3]:

Two or more weighting schemes on the datasets are tested with different imbalance ratios, that determine the better performance of weighted ELM.

Addressing the class imbalance problem in medical datasets [4]:

To reduce the ratio gap between the majority classes with minority class the algorithm is identified. In order to understand the quality of the training set, the minority samples were separated into three clusters and then grouped in various combination with the clusters of majority class samples**.**

Improved response modeling based on clustering, under-sampling, and ensemble [5]:

It mainly focuses on identifying the buyers, up-lift modeling finds the customers who will buy the product only when they are targeted by the marketing campaign. The improved response model by increasing response rate as well as reducing performance variation.

Inverse random under sampling for class imbalance problem and its application to multi-label classification [6]:

Under-sampling the majority classes create a large number of distinct training sets. For each training set it finds a decision boundary that separates the minority class from the majority class. By combining the multiple designs through fusion, construct a composite boundary between the majority class and the minority class.

Strategies for learning in class imbalance problems [7]:

Duplicating the minority class to eliminate imbalance, the TS (training set) does not add new information to the system. Moreover, working in that direction means to worsen the known computational burden of some learning algorithms, such as the NN (nearest neighbor) rule and the Multi-Layer Perceptron.

A learning method for the class imbalance problem with medical datasets [8]:

In the minority class, they creating ‘‘synthetic’’ samples rather than simply duplicating them, and their approach is known as SMOTE (synthetic minority over-sampling technique). In SMOTE, the synthetic examples are achieved in a less application-specific manner, that operates virtual sample generation in a ‘‘feature space’’ instead of a ‘‘data space’’.

New approach with ensemble method to address class imbalance problem [1]:

Rotation Forest promotes the accuracy of classifier by concentrating on class sample and SMOTE algorithm boost the performance of classifier in the minority class samples. It should be noted that although SMOTE-RO FO is less varied in general but has high accurate value and less error rate.

2.2 METHODOLOGY:

Class imbalance problem has been reorganized to be existing in lots of application domains, such as spotting unreliable telecommunication customers, detection of oil spills in satellite radar images, learning word pronunciations, text classifications, risk management, information retrieval and filtering tasks, medical diagnosis (e.g. rare disease and rare genes mutations),network monitoring and intrusion detection, fraud detection, shuttle system failure, earthquakes and nuclear explosions and helicopter gear-box fault monitoring.

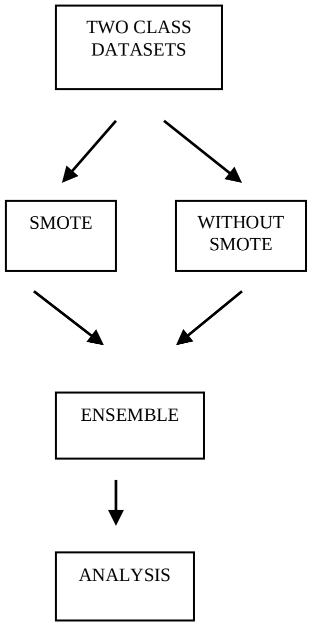
From the view of applications, the nature of the imbalance falls in two cases: The data are naturally imbalanced (e.g. credit card frauds and rare disease) or, the data are not naturally imbalanced but it is too expensive to obtain data of the minority class (e.g. shuttle failure) for learning. By using existing algorithms user cannot get good accuracy for all types of datasets. Although there are no consistent conclusions on which once are better recent studies suggest combining multiple classifiers.

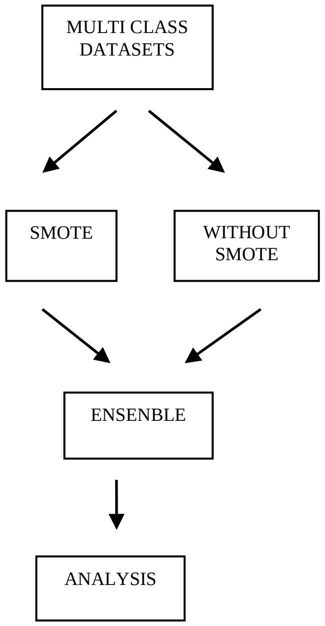
3. PROBLEM DEFINITION

The class is imbalanced when the number of records in one class is more than another one. The classification of this imbalanced class causes imbalanced distribution and poor predictive classification accuracy. Because of the imbalance class the accuracy is affected.

4. SYSTEM ARCHITECTURE

The accuracy can be predicted with the help of preprocessing method called SMOTE for two class datasets and multi class datasets. After preprocessing the datasets, the ensemble method is used to analyze the accuracy.





System architecture for two class imbalance datasets

System architecture for multi class imbalance datasets

5. IMPLEMENTATION

The accuracy is calculated using the WEKA tool and the version is WEKA 3.6.1. It is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from our own Java code. The accuracy is predicted using ensemble methods such as adaboost, Bagging, random subspace, rotation forest with different percentage split.

5.1 DATASET COLLECTION:

The datasets are collected from UCI repository, WEKA tool,etc.

Table 1 Dataset collections

Dataset	Instances	Class	Imbalanced Ratio
Vote	435	267,168	61%,39%
Supermarket	4627	2948,1679	63%,37%
Segment challenges	1500	205,220,208,220 204,236,207	13.67%,14.67%, 13.89%,14.67%, 13.67%,15.7%,13.8%
Credit	1000	700,300	70%,30%
Breast cancer	286	201,85	70%,30%

Diabetes	768	500,268	65%,35%
Ionosphere	351	126,225	36%,64%
Labor	57	20,37	35%,65%
Contact lenses	24	5,4,15	21%,17%,62.5%
Weather	14	9,5	64.3%,35.7%

5.2 IMBALANCED RATIO BEFORE AND AFTER PREPROCESSING:

The imbalance ratio of the different datasets before and after preprocessing can be obtained.

TABLE 2 Imbalanced ratio before and after preprocessing

Dataset	Before Preprocessing	After Preprocessing	Imbalance Ratio(After SMOTE)
Vote	435	603	44%,56%
Super market	4627	6306	47%,53%
Segment	1500	1500	13.67%, 14.67%, 13.89%, 14.67%, 13.69%, 15.7%, 13.8%
Credit	1000	2600	54%,46%
Cancer	286	371	54%,46%
Diabetes	768	1036	48%,52%
Iono sphere	351	225	53%,47%
Labor	57	77	52%,48%
Contact	24	51	20.8%, 16.7%, 62.5%
Weather	14	19	47.4%, 52.6%

5.3 ANALYSIS

CREDIT:

When applying preprocessing technique to the credit dataset, the accuracy of minority class is better when compared to accuracy before preprocessing. In random subspace the percentage split 66% as got greater accuracy while preprocessing in second class accuracy.

TABLE 3 First class accuracy

	Class=700(good),300(bad) Imbalanced ratio:70%(good),30%(bad)			Class=1400(good),1200(bad) Imbalanced ratio: 54%(good),46%(bad)		
	Instances=1000			Instances= 2600		
ALGORITHM	**Without SMOTE**			**With SMOTE**		
	50%	**66%**	**90%**	**50%**	**66%**	**90%**
ADABOOST(J48)	83.1	91.2	86.5	80	84.4	80.1
BAGGING(J48)	87.4	86	83.8	82.7	84.4	86
RANDOM SUBSPACE(J48)	92.3	97.2	89.2	87.8	88.8	89
ROTATION FOREST(J48)	82.8	86	87.8	85.9	85.7	86

TABLE 4 second class accuracy

ALGORITHM	Without SMOTE			With SMOTE		
	50%	**66%**	**90%**	**50%**	**66%**	**90%**
ADABOOST(J48)	41.8	34.4	38.5	74	86.4	79
BAGGING(J48)	41	46.7	42.1	83.8	86.4	86.3
RANDOM SUBSPACE(J48)	32.8	14.4	30.8	76.9	83.6	84.7
ROTATION FOREST(J48)	44	51.1	76.9	85	85.4	86.3

TABLE 5 Overall class accuracy

ALGORITHM	Without SMOTE			With SMOTE		
	50%	**66%**	**90%**	**50%**	**66%**	**90%**
ADABOOST(J48)	41.8	34.4	38.5	74	86.4	79
BAGGING(J48)	41	46.7	42.1	83.8	86.4	86.3
RANDOM SUBSPACE(J48)	32.8	14.4	30.8	76.9	83.6	84.7
ROTATION FOREST(J48)	44	51.1	76.9	85	85.4	86.3

DIABETES:

When applying preprocessing technique to the diabetes dataset, the overall accuracy predicted is better when compared to accuracy before preprocessing. In random subspace the percentage split 66% as got greater accuracy while preprocessing in second class accuracy.

TABLE 6 First class accuracy

	Class=500(tested negative),268(tested positive) **Imbalanced ratio:65%(tested negative),35%(tested positive)**			**Class=500(tested negative),536(tested positive)** **Imbalanced ratio: 48%(tested negative),52%(tested positive)**		
	Instances=768			**Instances= 1036**		
ALGORITHM	**Without SMOTE**			**With SMOTE**		
	50%	**66%**	**90%**	**50%**	**66%**	**90%**
ADABOOST(J48)	89	83.1	82.7	75	70.7	78.7
BAGGING(J48)	84.7	87.1	82.1	75	70.7	78.7
RANDOM SUBSPACE(J48)	87.5	92.1	92.3	72.3	71	80.9
ROTATION FOREST(J48)	86.7	86.5	88.5	74.6	71.3	83

TABLE 7 Second class accuracy

ALGORITHM	Without SMOTE			With SMOTE		
	50%	**66%**	**90%**	**50%**	**66%**	**90%**
ADABOOST(J48)	55.9	66.3	60	72.1	77	82.5
BAGGING(J48)	58.9	60.2	68	82.4	79.2	84.2
RANDOM SUBSPACE(J48)	48.1	44.6	48	80.9	83.7	89.5
ROTATION FOREST(J48)	57.4	62.7	64	79.4	83.7	94.7

TABLE 8 Overall class accuracy

ALGORITHM	Without SMOTE			With SMOTE		
	50%	66%	90%	50%	66%	90%
ADABOOST(J48)	77.9	77.8	75.3	73.6	73.9	80.8
BAGGING(J48)	76	78.5	77.9	78.8	75	81.7
RANDOM SUBSPACE(J48)	74.2	77	77.9	76.6	77.6	85.6
ROTATION FOREST(J48)	76.8	78.9	80.5	77	77.6	89.4

BREAST CANCER:

When applying preprocessing technique to the breast cancer dataset, the second class accuracy is predicted better when compared to accuracy before preprocessing. In second class accuracy the random subspace for the percentage.

TABLE 9 First class accuracy

	Class=201(no recurrence events),85(recurrence events) Imbalanced ratio:70%(no recurrence events),30%(recurrence events)			Class=201(no recurrence events),170(recurrence events) Imbalanced ratio:54%(no recurrence events),46%(recurrence events)		
	Instances=286			Instances= 371		
ALGORITHM	Without SMOTE			With SMOTE		
	50%	66%	90%	50%	66%	90%
ADABOOST(J48)	94.8	89.1	89.5	67	69.7	60
BAGGING(J48)	96.9	95.3	94.7	61.2	66.7	65
RANDOM SUBSPACE(J48)	100	92.1	94.7	77.3	71	75
ROTATION FOREST(J48)	87.5	90.6	89.5	68.9	74.2	80

TABLE 10 Second class accuracy

ALGORITHM	Without SMOTE			With SMOTE		
	50%	66%	90%	50%	66%	90%
ADABOOST(J48)	34	39.4	20	65.9	68.3	64.7
BAGGING(J48)	10.6	24.2	20	65.9	65	76.5
RANDOM SUBSPACE(J48)	0	24.2	10	73.3	83.7	64.7
ROTATION FOREST(J48)	34	30.3	20	68.3	65	76.5

TABLE 11 Overall class accuracy

ALGORITHM	Without SMOTE			With SMOTE		
	50%	66%	90%	50%	66%	90%
ADABOOST(J48)	74.8	72.2	69	66.5	69	62.2
BAGGING(J48)	68.5	71.1	77.9	63.2	65.9	70.3
RANDOM SUBSPACE(J48)	67.1	77.1	65.5	75.4	77.6	70.3
ROTATION FOREST(J48)	69.9	70.1	65.5	68.6	69.8	78.4

6. RESULT & DISCUSSIONS

Four ensemble approaches (Random subspace, bagging, rotation forest and adaboost) are considered to analyze the best performer for the class imbalanced problem. Here 10 class imbalanced datasets are considered. From the above analysis, it is clear that rotation forest and adaboost outperforms other ensemble approaches like random subspace and bagging in many occasions. Similarly, accuracy of the minority class also improved.

7. CONCLUSION

Many domains are affected by the problem of class imbalance. That is, when instances of one class greatly out number instances of the other class (es), it can challenge to construct classifier that effectively identifies the instances of underrepresented class. Several techniques have been proposed for dealing with problem of class imbalance. In this paper performance of existing algorithms are compared and analyzed. The performance shows that ensemble classifiers rotation forest gives the best accuracy when compared to AdaBoost, random subspace, bagging algorithms.

REFERENCES

[1] Barandelaa.R, Sanchezb.J.S, Garca.V, Rangela.E, 2011 . Strategies for learning in class imbalance problems. *Elsevier Science Ltd*.

[2] Der-Chiang Li ,Chiao-WenLiu ,SusanC.Hub, 2011. A learning method for the class imbalance problem with medical datasets. *Elsevier Science Ltd*.

[3] Gopika.D, Azhagusundari.B, August 2014. A Novel Approach on Ensemble Classifiers with Fast Rotation Forest Algorithm. *International Journal of Innovative Research in Computer and Communication Engineering*, Vol. 2, Issue 8.

[4] Mostafizur Rahman.M and Davis.D.N. April 2013 Addressing the Class Imbalance Problem in Medical Datasets. *International Journal of Machine Learning and Computing*, Vol. 3, No. 2.

[5] Muhammad AtifTahir, JosefKittler, FeiYan, 2012. Inverse random under sampling for class imbalance problem and its application to multi-label classification. *Elsevier Science Ltd*.

[6] Pilsung Kang Sungzoon Cho Douglas L. MacLachlan., 2012. Improved response modeling based on clustering, under-sampling, and ensemble . *Elsevier Science Ltd*.

[7] Seyyedali Fattahi, Zalinda Othman, Zulaiha Ali Othman, 2015. New approach with ensemble method to address class imbalance problem. *Journal of Theoretical and Applied Information Technology*.

[8] Weiwei Zong, Guang-Bin Huang, Yiqiang Chen, 2013. Weighted extreme learning machine for imbalance learning. *Elsevier Science Ltd., Neuro computing* 101 229–242,

[9] https://en.wikipedia.org/wiki/Weka/

[10] https://en.wikipedia.org/wiki/Data_mining

[11] http://www.cs.waikato.ac.nz/ml/weka/

[12] http://archive.ics.uci.edu/ml/