

Enhanced AdaBoost Algorithm with Modified Weighting Scheme for Imbalanced Problems

E. Prabhakar*

*Assistant Professor, Department of Computer Science & Engineering, Nandha College of Technology, Erode, Tamilnadu, INDIA.
E-Mail: prabhakarit10 {at} gmail {dot} com

Abstract— Class imbalance is notable critical issues arise during classification of particular applications datasets which is real world. This would leads to added misclassification in specifically one class that compares to other classes. This issue called class imbalance would occurs if one class possess large amount of records and in other hand alternative class (es) possess specifically little number of the records. In order to deal with class imbalance problem, various approaches are introduced and carried out. These methods are primarily based on under sampling the records in majority class or over sampling the minority class records. However, both approaches will produce over fitting or under fitting problems for the trained classifier. This paper deals with the modification of adaboost to enhance the classification performance. Class reweight strategy in adaboost is modified based on percentage of misclassification in previous iteration.

Keywords—Class-Imbalance; Over-Sampling; ROS; RUS; SMOTE; Under-Sampling.

Abbreviations—Expansion Condensed Nearest Neighbor (CNN); Edited Nearest Neighbour Rule (ENN); Enhanced AdaBoost Algorithm with Modified Weighting Scheme (EAAMWS).

I. INTRODUCTION

DATA mining is one among the processes which extract the information from particular data set and further to transform that into an understandable form for future use. In most real world applications, the data sets could be imbalanced class distribution, which means most records are in particular majority class and small records in minority class. There are two kinds of imbalances in a data set. One is between-class imbalance; in this type some classes have large number of records than others. Another one is the within-class imbalance in which some subsets in one class comprises highly fewer instances than remaining subsets of same class. If we train the imbalanced data, then mostly records in the majority class will be trained heavily. If we balance data either by using over sampling or under sampling, examples in the both the class will be equally trained. Thus it considers as necessary for select the training records which are suitable for classification in imbalanced class distribution problem.

Classification is an approach for finding rules which partition data into various groups. Classification technique could consider the data which are distributed uniformly between the various classes. The objective is to use the training dataset to build a model of the class label based on the other attributes such that the model can be used to classify

new data. Due to availability of many datasets in real world applications with imbalanced class distribution issues, it is not much easier for getting higher accuracy. Thus authors step in to focusing on the imbalanced class distribution to increases the notable classification accuracy for imbalanced data.

Imbalanced data sets exists in many real-world domains, such as Fraud Detection, risk management, financial data analysis, telecommunication industry, Intrusion Detection and so on.

The complete overview of the class imbalance is provided in this section. The review work of this paper is provided in section II. Proposed research technique is provided in section III. The experimental results of the research technique are provided in section IV. The conclusion and future work of this paper is provided in section V.

II. RELATED WORKS

Imbalanced data set are made to be balanced either by using algorithm or data level based approach.

2.1. Data-Level Approaches

In data level, the aim is to re-balance class distribution by means of re-sampling the particular data space. The most important direct ways is to deal with the class imbalance is

altering the class distributions which are towards the more balanced distribution. Sampling approach could divide as over-sampling and under-sampling approach which are specified in [Qiong Gu et al., 2].

2.1.1. Under-Sampling

There are many samples of the one class than other in imbalanced class distribution problem, thus under-sampling mechanism would helpful for reducing samples there in majority class.

A. Random Under-Sampling

Simply RUS which is a simplest approach of the under-sampling and it would select the subset of that majority class by randomly. Finally that would combine with the minority class as one of training set.

B. Tomek Links

Tomek Links is used as under-sampling method which is also said to be as data cleaning method. As similar as under-sampling, only examples are belonging to majority class are gets eliminated. Additionally, as data cleaning approach, examples of both the classes are made to remove. The main disadvantage of Tomek Link under-sampling is, this approach could discard the useful data which are potentially and consider as important for induction process [Batista et al., 1].

C. Condensed Nearest Neighbor Rule (CNN)

The basic idea or knowledge behind this CNN implementation is of eliminating the examples from majority class which are found to be distant from decision border, since those examples are might consider as least relevant for specific learning [Batista et al., 1].

D. One-sided Selection

One-sided Selection is one of heuristic under-sampling mechanism that balances the data set by removing the redundant instances and noise of majority class [Batista et al., 1].

E. Neighborhood Cleaning Rule

Neighborhood Cleaning Rule shortly called ENN which uses Wilson's Edited Nearest Neighbour Rule (ENN) in the sense of removing the majority class instance. ENN would remove any of examples where the class label is differing from class in three nearest neighbours which are at least almost present. NCL would change the ENN in the sense of increasing data cleaning [Batista et al., 1].

F. Advantage and Disadvantage

Under-sampling could improve the information loss. On the other hand there is no specific over-fitting is found and not found as time consuming during process.

2.1.2. Over-Sampling

The over-sampling mechanism would enhance the amount of minority class examples by means of replicate the examples

in minority class or about synthetic generation of the examples to minimize the imbalanced distribution degree.

A. Random Over-Sampling

Simply ROS is a non-heuristic approach and it would replicate the examples of minority class by randomly by achieving the more balanced distribution.

B. SMOTE

Smote would add the new minority examples additionally by means of extrapolating between the preexisting those minority examples which are rather than the simple duplicating of original examples. Thus it creates and provides the synthetic minority class examples by means of selecting the nearest minority neighbours of particular minority example and that generates the new minority class examples along with lines and every nearest minority neighbour [Chawla et al., 5].

C. Borderline SMOTE

Simply BSM which selects the minority class examples that are assumed as on border of minority decision region in specific feature-space. Additionally the SMOTE has been performed to oversampling only those instances rather oversampling all or a random subset of them [Batista et al., 1].

D. Advantage and Disadvantage

Over-sampling would improve the likelihood of over-fitting which occurs and could take high time for over-sample due to the adding of samples. But no information loss is found.

2.1.3. Combination of Over-Sampling Method with the Under-Sampling Method

A. Smote + Tomek links

Even though the over-sampling of minority class examples are balance the class distributions, some other issues are present usually in using data sets with the skewed class distributions which are not found as solved. By frequently, the class clusters are not well described, since some of majority class examples are might invading that minority class space. The opposite also be true, since the interpolating minority class examples are expands the minority class clusters, which introduce the artificial minority class examples too which are deep in majority class space. Inducing specific classifier under that situation could leads to the over-fitting. In order to creating the defined class clusters that are better, the applying of Tomek links could oversampled the training set as data cleaning approach. Finally, instead of remove the majority class examples only which form the Tomek links, examples of both the classes are removing [Batista et al., 1].

B. Smote + ENN

Thus motivation behind this approach is same as Smote + Tomek links. In this ENN could remove the many examples

than Tomek links, and it has been expected as that would provides more in depth data cleaning. By differing from the NCL, that is an under-sampling approach, ENN is used for removing the examples of both classes. Thus, example which is misclassified by means of the three nearest neighbours is removing from training set [Batista et al., 1].

2.2. Algorithm Level Approaches

The approach at algorithm level would provide the steps for the classifier in the sense of classifying the examples. The standard boosting algorithm, for example Adaboost which enhances the votes or weights of the examples that are misclassified and in other side would minimize the weights for one correctly classified by means of same proportion. This is take place by without consider the imbalance of data sets [Seiffert et al., 6]. Due to this reason, the algorithms which are traditional boosting approach are found as not performing well on minority class. In those recent years, some of algorithms have been studied in order to combine data level techniques with algorithm level techniques in terms of solving the class imbalance problem [Chawla et al., 3].

2.2.1. Boosting

Boosting is a technique that used for improving the classification performance. While more data sampling approaches are gets designed by specifically to addressing class imbalance problem. Thus boosting technique is for increasing the performance in any of the weak classifier. The boosting algorithm which are widely using is AdaBoost that iteratively could builds an ensemble of models [Seiffert et al., 6].

2.2.2. AdaBoost

AdaBoost is an algorithm level approach, could not have any of pre-processing approach. It took much iteration, during an every iteration the example weights are changed with aim of classifying the examples correctly for next iteration, that are incorrectly classified in current iteration. Without balancing the specific imbalance problem that enhances the accuracy [Seiffert et al., 6].

2.2.3. SMOTEBOOST

SMOTEBOOST would combine the Smote approach with Boosting algorithm is depending on the AdaBoost. During every iteration of the boosting, SMOTE has been applied for training data which achieves the more balancing training data set [Seiffert et al., 6].

2.2.4. RUSBOOST

RUSBOOST improves the performance of specific trained data that combines AdaBoost algorithm with RUS algorithm. RUSBoost is depending on the Smote Boost which turns based on AdaBoost by means of over-sampling technique namely SMOTE may helps to balancing the class distribution. Whereas the AdaBoost increase the performance of classifier uses these balanced data. RUSBoost would

achieve similar result which uses the RUS rather than the SMOTE [Seiffert et al., 6].

III. PROPOSED WORK

There are various mechanisms are introduced could alleviate the class imbalance problem. This paper deals with the algorithm level approach to solve the class imbalance problem. Enhanced AdaBoost Algorithm with Modified Weighting Scheme (EAAMWS) is the modifications of class reweight strategy in adaboost algorithm. Class reweight for majority and minority class is applied differently in an every iteration based on the percentage of misclassification in previous iterations. It will be suitable for handle the class imbalance problem in two-class and multi-class data very well.

Let consider the two class imbalance problem. Class has a very high amount of records, thus it has been considered as MAC that is majority class. Class b consists of very little records and it considered as the minority class namely MIC.

3.1. Procedure

NMAC=Number of records taken for training from majority class in first iteration

NMIC=Number of records taken for training from minority class in first iteration

PCMAC=Percentage of records correctly classified which was taken for training from majority class in first iteration

PCMIC= Percentage of records correctly classified which was taken for training from minority class in first iteration

CRMIC=Class reweight in majority class

CRMIC=Class reweight in minority class

1. Apply any classification algorithm (C4.5 or J48) in first iteration
2. Calculate the number of correctly classified records in each class after first iteration
3. Now calculate the PCMAC and PCMIC
4. If PCMAC is greater than PCMIC then follow the below steps in second iteration
 $CRMIC=1$
 $CRMIC=PCMAC/PCMIC$
5. If PCMIC is greater than PCMAC then follow the below steps in second iteration
 $CRMIC=1$
 $CRMIC=PCMIC/PCMAC$
6. Repeat the above steps 4 or 5 in next iterations

IV. EXPERIMENTAL RESULTS

4.1. Datasets Features

Data sets were collected from UCI Repositories to calculate the performance of traditional algorithms. Data set contains

either two class imbalance or multi class imbalance. The information of the data sets is showed in below table.

Table I: Data Set Characteristics

Datasets	Instances	Attributes	Class
Mushroom	8124	22	2
Magic Gamma	19019	11	2
Balance Scale	625	5	3
Thyroid	721	22	3
Car Evaluation	1727	7	4
Vehicle	946	19	4

4.2. Performance Metrics

This paper uses the confusion matrix and two additional metrics which were derived from confusion matrix as the evaluation method to calculate the performance. True positive represents the positive records that are correctly classified while True negative represents the positive records that are incorrectly classified. True negative represents negative records that are correctly classified while false negative represents negative class that are incorrectly classified [Nguyen et al., 4].

Table 2: Confusion Matrix

Actual Class/ Predicted Class	Class1	Class2
Class1	True positive	False negative
Class2	False positive	True negative

Let,

TMAC = Total number of records in the majority class

TMIC = Total number of records in the minority class

TOC = Total number of records in both of class

CMAC = Total number of correctly classified records in the majority class

CMIC = Total number of correctly classified records in the minority class

COC = Total number of correctly classified records in the minority class

Accuracy has been calculated by means of formula,

Overall Accuracy (OAA) = COC / TOC

Majority Accuracy (MAA) = CMAC / TMAC

Minority Accuracy (MIA) = CMIC / TMIC

4.3. Result Analysis

Table 3: Two-Class Datasets

Datasets	Methods	MAA	MIA	OAA
Vote	Imbalanced	95.50	95.23	95.40
	ROS	95.88	96.62	96.25
	RUS	95.83	95.83	95.83
	SMOTE	99.25	95.88	97.76
Mush Room	Imbalanced	94.86	97.54	96.15
	ROS	99.26	97.90	98.58
	RUS	98.31	97.75	98.30
	SMOTE	99.27	98.25	99.36
Abalone	Imbalanced	99.57	16.66	95.49
	ROS	96.71	100	98.45
	RUS	95.15	35.29	90.98
	SMOTE	100	99.15	99.55

Magic Gamma	Imbalanced	80.11	74.35	78.09
	ROS	78.91	74.74	76.83
	RUS	78.60	75.91	77.25
	SMOTE	67.97	83.22	75.90
Labour	Imbalanced	80.0	91.89	87.71
	ROS	97.29	97.29	97.29
	RUS	95.0	70.0	82.5
	SMOTE	94.59	92.50	93.50
Ionosphere	Imbalanced	76.98	98.66	90.88
	ROS	80.44	92.0	86.22
	RUS	77.77	96.03	86.90
	SMOTE	93.77	79.36	86.16
Diabetes	Imbalanced	84.60	55.22	74.34
	ROS	76.6	71.6	74.1
	RUS	91.04	78.73	84.88
	SMOTE	77.99	72.60	75.38

Table 4: Three-Class Datasets

Datasets	Methods	C1A	C2A	C3A	OAA
Balance Scale	Imbalanced	7.14	87.09	81.6	77.8
	ROS	90.16	81.52	82.6	84.0
	RUS	15.79	92.47	81.2	80.2
	SMOTE	80.00	83.87	87.8	84.4
Nursery	Imbalanced	0.0	66.66	100	70.9
	ROS	77.21	21.62	47.6	41.6
	RUS	2.30	18.36	45.5	29.6
	SMOTE	75.17	0.0	89.3	52.1

Table 5: Four-Class Datasets

Datasets	Methods	C1A	C2A	C3A	C4A	OAA
Vehicle	Imbalanced	0.0	61.66	5.41	100	40.90
	ROS	0.0	69.58	0.0	99.5	42.29
	RUS	0.0	61.06	6.19	100	41.81
	SMOTE	0.0	67.45	4.36	99.5	42.85
Car Evaluation	Imbalanced	97.8	95.31	90.6	73.9	96.06
	ROS	100	100	97.5	0.0	99.38
	RUS	100	96.92	100	98.4	98.84
	SMOTE	87.8	0.0	100	0.0	57.86

Table 6: Segment- Seven Class Dataset

Methods	C1A	C2A	C3A	C4A	C5A	C6A	C7A	OAA
Imbalanced	0	100	0	0	0	100	0	30
ROS	0	100	0	0	99	98	0	53
RUS	100	0	0	0	37	89	0	43
SMOTE	0	100	0	0	0	88	0	30

Table 7: Glass – Seven Class Dataset

Methods	C1A	C2A	C3A	C4A	C5A	C6A	C7A	OAA
Imbalanced	100	0	0	0	0	0	90	45
ROS	100	0	0	93	0	0	0	47
RUS	0	100	0	0	0	0	88	44
SMOTE	100	0	0	0	100	0	0	39

Here,

C1A=Accuracy of class1

C2A=Accuracy of class2

C3A=Accuracy of class3

C4A=Accuracy of class4

C5A=Accuracy of class5

C6A=Accuracy of class6

C7A=Accuracy of class7

From the table 3, 4, 5, 6, 7, it is clear that most of the times traditional data mining algorithms like RUS, ROS and SMOTE are not suitable for all type datasets. Our algorithm will perform better than these algorithms, because it changes the class weight after an each iteration of classification.

V. CONCLUSION

The Traditional data mining techniques are focused on various pre-processing approaches like the under-sampling or over-sampling by solving class imbalance problem. Both the techniques performed well compared to the imbalanced data but we need to loss either some useful data or time when we use above techniques. In this paper class balance based on class reweight strategy in AdaBoost without any pre-processing methods. It performs better than traditional data mining algorithms. Because of it balance the class at the time of training.

REFERENCES

- [1] G.E.A.P.A. Batista, R.C. Prati & M.C. Monard (2004), "A Study of the Behaviour of Several Methods for Balancing Machine Learning Training Data", *ACM SIGKDD Explorations Newsletter - Special Issue on Learning from Imbalanced Datasets*, Vol. 6, No. 1, Pp. 20–29.
- [2] Qiong Gu, Zhihua Cai, Li Zhu & Bo Huang (2008), "Data Mining on Imbalanced Data Sets", *International Conference on Advanced Computer Theory and Engineering*.
- [3] N.V. Chawla, N. Japkowicz & A. Kolcz (2004), "Editorial: Special Issue on Learning from Imbalanced Data Sets", *Sigkdd Explorations*, Vol. 6, No. 1, Pp. 1–6.
- [4] G.H. Nguyen, A. Bouzerdoum & S.L. Phung (2009), "Learning Pattern Classification Tasks with Imbalanced Data Sets", *University of Wollongong*.
- [5] N.V. Chawla, K.W. Bowyer, L.O. Hall and W.P. Kegelmeyer (2002), "SMOTE: Synthetic Minority Over-Sampling Technique", *Journal of the Artificial Intelligence Research*, Vol. 16.
- [6] C. Seiffert, T.M. Khoshgoftaar, J.V. Hulse & A. Napolitano (2010), "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance", *IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems and Humans*, Vol. 40, No. 1, Pp. 185–197.