WIDS Approach for Class Imbalance Using Hybrid Sampling

D. Durga Prasad
PP COMP SCI ENGG 431, Dept. of CSE, Rayalaseema University, Kurnool(Dt), India

Dr. K.nageswara Rao
Principal, PSR & CMR College of Engineering and Technology, Vijayawada, A.P., India.

Abstract- The knowledge discovery from the imbalance datasets is a challenging task due to the intrinsic properties of the data. The existing algorithms are not efficient for deriving the hidden knowledge for imbalance datasets. In this paper we proposes a set of new hybrid framework using within classes instances oversampling in the minority subset and intelligent under sampling in the majority subset known as With In class Diverse Sampling (WIDS) for discovery of data using by the hybrid approach. It overcomes the disadvantage of weak unbalanced data distribution. Further the proposed framework eliminates the requirement of repeated oversampling weak instances in minority subset by repeatedly removing of useless instances in the oversampling approaches for balancing the data. The WIDS algorithm is compared with SMOTE approach on 15 imbalance datasets from UCI Repository. The results suggest that the proposed approach is efficient than the compared approach in terms of AUC, Precision, Recall and F-measure.

Keywords – Classification, Imbalanced data, Under sampling, Over Sampling, Hybrid Sampling and WIDS.

I. INTRODUCTION

One of the most popular techniques for alleviating the problems associated with class imbalance is data sampling. Data sampling alters the distribution of the training data to achieve a more balanced training data set. The data sets are classified in two categories those are: Under or Over sampling of classes. Under sampling removes majority class examples from the training data, while oversampling adds examples to the minority class. Under Sampling and Over Sampling methods are accomplished may be in randomly or intelligently techniques.

The random sampling techniques either duplicate (oversampling) or remove (under sampling) random examples from the training data. Synthetic minority oversampling technique (SMOTE) [1] is a more intelligent oversampling technique that creates new minority class examples, rather than duplicating existing ones. Wilson's editing (WE) intelligently under samples data by only removing examples that are thought to be noisy. In this study, we investigate the impact of intelligent under sampling technique on the performance of the classification algorithms. While the impacts of noise and imbalance have been frequently investigated in isolation, their combined impacts have not received enough attention in research, particularly with respect to classification algorithms. To alleviate this deficiency, we present a comprehensive empirical investigation of learning from noisy and imbalanced data using classification techniques.

The present paper proposes a novel hybrid framework using diverse (under sampling and oversampling) techniques. There are many Class Imbalance Learning (CIL) methods based on the hybrid approaches. So far very few researchers have attempted to use Hybrid approaches with both under and oversampling.

II. RELATED WORK

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

Class Imbalance Learning Using Intelligent Sampling:

This chapter mainly focuses on exploration of WIDS model by using experimentation using different decision tree models. The main linear regression models used for experimentation in this section are Logistic Regression and Support Vector Machines. In the next subsection a brief description about each decision tree model is given.

2.1 Within class Imbalance Majority Under Sampling (WIMUS) Technique:

The proposed WIMUS [2] method consists of two stages. In the first stage the majority and minority instances are divided and within class are identified in majority subset for under sampling noisy and border line instances, thereby

forming new improved dataset. In the second stage the improved dataset is applied to a base algorithm for evaluating the performance. Here, Random Forest is considered as the base algorithm.

The working style of under-sampling tries to decrease the number of weak or noise examples. In this, the weak instances features are to be deleted and which area identified according to a well-established filter and intelligent technique. The number of instances eliminated will belong to the 'k' feature selected by filter and intelligent technique. The above said routine is employed, which removes examples suffering from feature, at first to remove class label noises and also remove borderline and its category that is outliers. Feature to Class label noises are the examples whose influence is not seen for the decision of the class for that particular feature. Here, they are identified by the limited range categories, using the above said technique. In detail, at first some examples are deleted temporary from Nstrong, a new dataset created with strong instances. Then, for a class to be shrank, all its examples inside of Nstrong are classified. If the classification is correct, and the accuracy is increased then the examples deleted temporary are regarded as being feature class label noises. Borderline examples are close to the boundaries between different classes for a specific feature. They are unreliable because even a small amount of attribute noise can send the example to the wrong side of the boundary. The outliers are the examples which are very rare from the remaining set of examples. These are examples are of very rare use to the classification and thus to be removed for better performance.

2.2 Within class Minority Oversampling TEchnique (WIMOTE):

In the proposed WIMOTE [3] technique, the most important within class sub classes are identified in the minority subset and those instances are recursively oversampled to improve class imbalance learning. This oversampling Class Imbalance Learning (CIL) approach overcomes the weakness of resampling noisy and less priority instances.

In over sampling, we will take minority data subset for further visualization analysis to identify within class imbalances. Minority subset can be further analyzed to find the noisy or borderline instances so that we can eliminate those. The influencing attributes or characteristics of the weak attributes and delete the noisy or weak instances relating to that feature. How to choose the noisy instances relating to that within classes from the dataset set? for this we find a limit with the number of samples are less to specify a simple hint that those instances coming in that range or very rare or noise. We will intelligently detect and remove those instances which are in narrow ranges of that particular within classes. This process can be applied on all the within class imbalances identified for each dataset.

The oversampling of the instances can be done efficiently, if the weightages are assigned to the particular within class imbalances in the minority subset. In this stage, the weightages are assigned to within class imbalances depending upon the density of the within class imbalances. The more dense the within classes, more the weightages of that within classes and vice versa.

III. PROPOSED WIDS APPROACH FRAMEWORK

The different components of our new proposed WIDS framework are elaborated in the next subsections.

Phase I: Preparation of the Majority and Minority subsets

The datasets is partitioned into majority and minority subsets. As we are concentrating on both over sampling and under-sampling, we will take minority and majority data subset for further visualization analysis to identify within class imbalances.

Phase II: Improve within class imbalances by removing noisy and borderline instances

Minority subset can be further analyzed to find the noisy or borderline instances so that we can eliminate those. For finding the weak instances one of the ways is that find most influencing attributes or features and then remove ranges of the noisy or weak attributes relating to that feature.

How to choose the noisy instances relating to that within class imbalances from the dataset set? We can find a range where the number of samples are less can give you a simple hint that those instances coming in that range or very rare or noise. We will intelligently detect and remove those instances which are in narrow ranges of that particular within class imbalances. This process can be applied on all the within class imbalances identified for each dataset.

Phase III: Applying under sampling on within class imbalances

Apply WIMUS algorithm for under-sampling the instances from the majority subset. In WIMUS the weak instances related to the specific features are to be eliminated, which is identified according to a well-established filter and intelligent technique. The number of instances eliminated will belong to the *'k'* feature selected by filter and intelligent technique. Here, the above said routine is employed, which removes examples suffering from feature to class label noises at first and then removes borderline examples and examples of outlier category. Feature to Class label noises are the examples whose influence is not seen for the decision of the class for that particular feature. Here, they are identified by the limited range categories, using the above said technique.

Phase IV: Applying oversampling on within class imbalances

The oversampling of the instances can be done on the improved within class imbalances produced in the earlier phase. The oversampling can be done as follows:

Apply resampling supervised filter on the within classes for generating synthetic instances. The synthetic minority instances generated can have a percentage of instances which can be replica of the pure instances and reaming percentage of instances are of the hybrid quality of synthetic instances generated by combing two or more instances from the pure minority subset. Perform oversampling on within classes can help so as to form strong, efficient and more valuable rules for proper knowledge discovery.

Phase V: Forming the strong dataset

The minority subset and majority subset is combined to form a strong and almost balance dataset, which is used for learning on a base algorithm. In this case we have used random forest as the base algorithm. The proposed WIDS approach algorithm is summarized as below.

Algorithm: WIDS APPROACH

Input: A set of major subclass examples *P*, a set of minor subclass Examples *N*, *jPj* < *jNj*, and *Fj*, the feature set, j > 0.

Output: Average Measure {AUC, Precision, F-Measure, TP Rate, TN Rate}

Phase I: Initial Phase:

1: begin
2: $k \leftarrow 1$, j$\leftarrow$1.
3: Apply Visualization Technique on subset *N*,
4: Identify within classes *Cj from* N, j= number of within classes identified in visualization
5: repeat
6: k=k+1
7: Identify and remove the borderline and outlier instances for the within classes *Cj*.
8: until k = j

Phase II: Under sampling Phase

9: Apply Under-sampling on *Cj* within classes from P,
10: repeat
11: k=k+1
12: Remove '*Cj* × s' noisy, borderline instances from the majority examples in each within classes *Cj*.
13: until k = j

Phase III: Over sampling Phase

14: Apply Oversampling on *Cj* within classes from N,
15: repeat
16: k=k+1
17: Generate '*Cj* × s' synthetic positive examples from the minority examples in each within classes *Cj*.
18: until k = j

Phase IV: Validating Phase

19: Train and Learn A Base Classifier (random forest) using Improved *P* and *N*
20: end

IV. DATASETS FOR WIDS

Experiments are conducted using fifteen datasets from UCI [6] 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, and 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.	Breast	286	9	2.37
2.	Breast-cancer-w	699	9	1.90
3.	Horse-colic	368	22	1.71
4.	Credit-g	1,000	20	2.33
5.	Pima diabetes	768	8	1.87
6.	Heart-c	303	13	1.19
7.	Heart-h	294	13	1.77
8.	Heart-statlog	270	14	1.25
9.	Hepatitis	155	20	3.85
10.	Ionosphere	351	35	1.79
11.	Kr-vs-kp	3196	37	1.09
12.	Labor	57	17	1.85
13.	Mushroom	8124	23	1.08
14.	Sick	3772	30	15.32
15.	Sonar	208	13	1.15

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.

Tables 2-3 presents the performance of SMOTE and WIMUS methods averaged across all data sets. These tables give a general view of the performance of both SMOTE and WIMUS method using each of the four performance metrics.

$$F-measure = \frac{2 \times \Pr ecision \times \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \quad \text{---------- (1)}$$

The Precision measure is computed by,

$$\Pr ecision = \frac{TP}{(TP)+(FP)} \quad \text{----------(2)}$$

The Recall measure is computed by,

$$\operatorname{Re} call = \frac{TP}{(TP)+(FN)} \quad \text{---------- (3)}$$

The Area under Curve (AUC) [8] measure is computed by,

The F-measure Value is computed by,

Tl $AUC = \frac{TP_{RATE} + TN_{RATE}}{2}$ y, ---------- (4)

$$AUC = \frac{1 + TP_{RATE} - FP_{RATE}}{2} \quad \text{---------- (5)}$$

V. EXPERIMENTAL RESULTS

We have analyzed the performance of our proposed algorithm WIDS on class imbalance problem on the following twelve real-world datasets. The results of the tenfold cross validation with standard deviation are shown in Tables 2 &3. In Tables 2 &3, we can observe the results of our proposed algorithm WIDS Vs SMOTE algorithms with respect to AUC, Recall, Precision and F-measure.

We present the results of the comparison between WIDS and SMOTE. From these results we can make several observations. The developed WIDS based on diverse (under sampling and oversampling) technique generally given competitive results for SMOTE; the advantage of our methods is most visible in the breast_w, diabetes, labor, ionosphere and sick datasets. Finally, the method that most often registered wins is WIDS approach.

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

Datasets	SMOTE	WIDS
Breast	0.717±0.084●	0.976±0.021
Breast_w	0.967±0.025●	0.998±0.004
Colic	0.908±0.040●	0.985±0.013
Crerdit-g	0.778±0.041●	0.996±0.006
Diabetes	0.791±0.041●	0.991±0.008
Hepatitis	0.798±0.112●	0.986±0.029
Ionosphere	0.904±0.053●	1.000±0.001
Kr-vs-kp	0.999±0.001●	1.000±0.001
Labor	0.833±0.127●	0.993±0.024
Mushroom	1.000±0.00	1.000±0.000
Sick	0.962±0.025●	1.000±0.000
Sonar	0.814±0.090●	0.995±0.010

● Bold dot indicates the win of Proposed WIDS approach;

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

Datasets	SMOTE	WIDS
Breast	0.710±0.075●	0.951±0.032
Breast_w	0.974±0.025●	0.995±0.008
Colic	0.853±0.057●	0.958±0.026
Crerdit-g	0.768±0.034●	0.985±0.014
Diabetes	0.781±0.064●	0.977±0.018
Hepatitis	0.709±0.165●	0.960±0.081
Ionosphere	0.934±0.049●	0.992±0.017
Kr-vs-kp	0.996±0.005●	0.998±0.003
Labor	0.871±0.151●	0.974±0.081
Mushroom	1.000±0.000	1.000±0.000
Sick	0.983±0.007●	0.998±0.002
Sonar	0.863±0.068●	0.982±0.036

● Bold dot indicates the win of Proposed WIDS approach

REFERENCES:

[1] Imbalance", I. King et al. (Eds.): ICONIP 2006, Part II, LNCS 4233, pp. 21–30, 2006.Springer-Verlag Berlin Heidelberg 2006.

[2] GIOVANNA MENARDI, NICOLA TORELLI, "Training and assessing classification rules with unbalanced data", DEAMS working paper 2/2010.

[3] 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

[4] 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

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

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

[7] D. Durga Prasad and Dr K. Nagswara Rao (2016), "An Improved approach on Class Imbalance data using Within-Class Minority Oversampling Technique ", International Journal of Latest Trends in Engineering and Technology Vol.(7)Issue(4), pp.156-164.