Classifying Muti-Class Imbalance Data

[1]Marwa F. Al-Roby, [2]Alaa M. El-Halees
Computer Science Department, Islamic University of Gaza – Palestine
m87alroby@gmail..com , ALHALEES@iugaza.edu.ps

Abstract

Class imbalance is one of the challenging problems for data mining and machine learning techniques. The data in real-world applications often has imbalanced class distribution. That is occur when most examples are belong to a majority class and few example belong to a minority class. In this case, standard classifiers tend to classify all examples as a majority class and completely ignore the minority class. For this problem, researchers proposed some solutions at both data and algorithmic levels. Most efforts concentrate on binary class problems. However, binary class is not the only scenario where the class imbalance problem prevails. In the case of multi-class data sets, it is much more difficult to define the majority and minority classes. Hence, multi class classification in imbalanced data sets remains an important topic of research. In our research, we proposed new approach based on SOMTE (Synthetic Minority Over-sampling TEchnique) and clustering which is able to deal with imbalanced data problem involving multiple classes. We implemented our approach and experimental results show our approach is effective to deal with the multi class imbalanced data sets, and can improve the classification performance of minority class and its performance on the whole data set. In the best case, our F-measure improved from 66.91 to 95.18.

Keywords: *Data mining, Classification, Multi class classification, Class imbalanced problem, sampling methods.*

1. Introduction

The classification techniques usually assume a balanced class distribution (i.e. there data in the class is equally distributed). A classifier performs well when the classification technique is applied to a dataset evenly distributed among different classes. But many real applications face the imbalanced class distribution problem [9]. In this situation, the classification task imposes difficulties when the classes present in the training data are imbalanced. The imbalanced class distribution problem occurs when one class is represented by a large number of examples (majority class) while the other is represented by only a few (minority class). In this case, a classifier usually tends to predict that samples have the majority class and completely ignore the minority class. This is known as the class imbalance problem [8].

The imbalanced data problem can appear in two different types of data sets: binary problems, where one of the two classes comprises considerably more samples than the other and multi-class problems, where the applications have more than two classes and unbalanced class distribution hinder the classification performance. Most research efforts on imbalanced

data sets have traditionally concentrated on two-class problems. However, this is not the only scenario where the class imbalance problem prevails. In the case of multi-class data sets, it is much more difficult to define the majority and minority classes [5, 10]. Hence, multi class classification in imbalanced data sets remains an important topic of research. In our research we focused mainly onmulti class imbalance problem which the two-class problem is considered as a special case from multi-class problem.We depend on F-measure as a measure for classification for imbalanced data.

Some of researches have been done for imbalanced class distribution problem domain. These approaches have been introduced at both algorithm and data levels. At the data level, the objective is to re-balance the class distribution by re-sampling the data space including over sampling instances of the positive class and under sampling instances of the negative class [3, 4, 7, 8, and 12]. At algorithm level, solutions try to adapt the existing classifier learning algorithm to bias towards the positive class [3, 4, 7, 8, and 12].

The rest of the research is organized as follows: Section 2 for related work. Section 3 include the methodology and proposed model architecture. In Section 4, we discussthe experimental results and analysis. Section 5 draws the conclusion and summarizes the research achievement and future direction.

2. Related Work

In this section we review the most important ones in both data and algorithmic levels,The following are some well known works on imbalanced data mining implemented at data level:**Yen and Lee in [8]** proposed cluster-based under-sampling approaches for selecting the representative data as training data. In general the F-measure term does not exceed 79% which is considering low. Also, as stated before under sampling approach may lose useful information about the majority class.**Chawla et al. in [6]** presented theSynthetic Minority Over-sampling TEchnique (SMOTE) approach, which is generate synthetic minority samples by interpolating between two minority samples that lie together at an over sampling rate. The method is evaluated using the area under the Receiver Operating Characteristic curve (AUC) and ROC convex hull strategy. This method has been validated to be effective. However, they apply only for binary class.

At algorithm level:**Murphey et al. in [11]** proposed a new pattern classification algorithm, One Against Higher Order (OAHO), that effectively learn multi-class patterns from the imbalanced data. The idea is building a hierarchy of classifiers based on the data distribution. OAHO constructs K - 1 classifiers for K classes in a list of {C1;C2; :::;CK}. Although OAHO has been proposed to handle the imbalanced problem for multi-class classification, its performance is sensitive by the classifier order, as misclassification made by the top classifiers cannot be corrected by the lower classifier.**Adam et al. in [1]** solved imbalanced data set problem through introduced feed forward ANN that is used particle swarm optimization (PSO).PSO is an advancedoptimization intelligent technique that easy to implement inoptimization problems and it has been successfully appliedin various fields.We note they solved imbalanced problem at algorithm level with modify ANN, but we think solved imbalanced problem at data level is better because after that we can use this data with different classifier. For this reason we work at data level to handle imbalanced class distribution.

From the previous works we can conclude that we note few of works proposed for multi class problem because it is much more difficult to define the majority and minority classes. We preferred to work at the data level than work at the algorithm level because at data level after preprocessing data we can use this data with different classifier but at algorithm level we need to modify each classifier that is used with imbalanced datasets. Also we note in the researches which is used clustering technique, they determine the number of cluster manually. However, in our research we try to test the results with clustering data automatically. Also we think the performance of minority class can be improved.

3. Methodology

Our main objective in this research is increase the classification accuracy of minority class by avoiding the drawbacks of the existing methods. For that, we propose an efficient approach combine between both Synthetic Minority Over-sampling TEchnique (SOMTE) approach [6] and clustering approach which is able to deal with multi class imbalanced data problem. To do that we propose the following steps in the preprocessing stage which are:

1. Clustering the data into clusters using random clustering algorithm to obtain clusters with equal number of instance approximately.
2. Also in other experiments we used X-mean algorithm to test the effect of determine the number of clusters automatically.X-mean is K-mean extended by an improve structure part through efficient estimation of the number of cluster automatically [2]. That means we do not need to enter the number of clusters by ourselves. The x-mean algorithm starts with K (k: number of cluster) equal to the lower bound of the given range and continues to add centroids where they are needed until the upper bound is reached. During this process, the centroid set that achieves the best score is recorded, and this is the one that is finally output.
3. Use over sampling which duplicates the sample of the minority class and adding them to data set.
4. Use SMOTE approach which generates new synthetic minority instances by interpolating between several minority examples that lie close together. SMOTE was introduced by Cieslak and Chawla [6], who suggested a local implementation of sampling based on create "synthetic" instances from existing minority class samples.

4. Experimental Results and Analysis

In this section, we present and analyze experimental results. We used different machine learning classifier for our experiments named, rule induction, naïve Bayes, decision tree and neural network on the selected datasets to classify the instances. For evaluation purpose, we use 10 cross-validation method. Also we assume that the ratio of the number of majority class samples to the number of minority class samples in the training data is set to be 1:1. In other word, there are the whole 100 majority class samples and there are must existing 100 minority class samples in this training data set. Six data sets are chosen from different real domain, characteristics and sizes. Five from data sets (page blocks, cardiotocography, car evaluation,

auto MPG and glass identification) represent multi class problem case and the other one data set (breast cancer-w) represent two class problem case. General information about these eight data sets is tabulated in Table (1).

Table 1: Summary of data sets

Data set	Data type	# instance	# Attribute	# class	Class distribution	% Class distribution	Reference
Page Blocks	Real	5473	10	5	• Text: 4913 • Horiz-line: 329 • Picture: 115 • Vrt-line: 88 • Graphic: 28	Text:89.8% Horiz-line:6% Picture:2% Vrt-line:1.6% Graphic:0.6%	[13]
Cardiotocography	Real	2126	23	3	• Normal: 1655 • Suspect: 295 • Pathology: 176	Normal:77.8% Suspect:13.9% Pathology:8.3%	[14]
Car Evaluation	Categorical	1728	6	4	•Unacc: 1210 •Acc: 384 •Good: 69 •Vgood: 65	Unacc:70% Acc:22% Good:4% Vgood:4%	[15]
Auto MPG	Real	398	8	5	•Class 4: 204 •Class 8: 103 •Class 6: 84 •Class 3: 4 •Class 5: 3	Class 4: 51.3% Class 8: 25.7% Class 6: 21% Class 3: 1% Class 5: 1%	[16]
Glass Identification	Real	214	10	6	•Class 2: 76 •Class 1: 70 •Class 7: 29 •Class 3: 17 •Class 5: 13 •Class 6: 9	Class 2: 35.5% Class 1: 32.7% Class 7:13.55% Class 3: 7.9% Class 5: 6% Class 6: 4%	[17]
Breast Cancer - w	Real	699	10	2	•Benign: 458 •Malignant: 241	Benign: 65.5% Malignant:34.5%	[18]

We can summarize our experiments results as is in rule induction, the highest F-measure result (92.07) was in our approach (SOMTE based on clustering). In naïve Bayes, the highest F-measure result (86.02) was in our approach. In decision tree, the highest F-measure result (95.18) was in our approach. In neural network, the highest F-measure result (95.64) was in over sample approach with three clusters. Fig.1. shows an overview of the all experiment results.

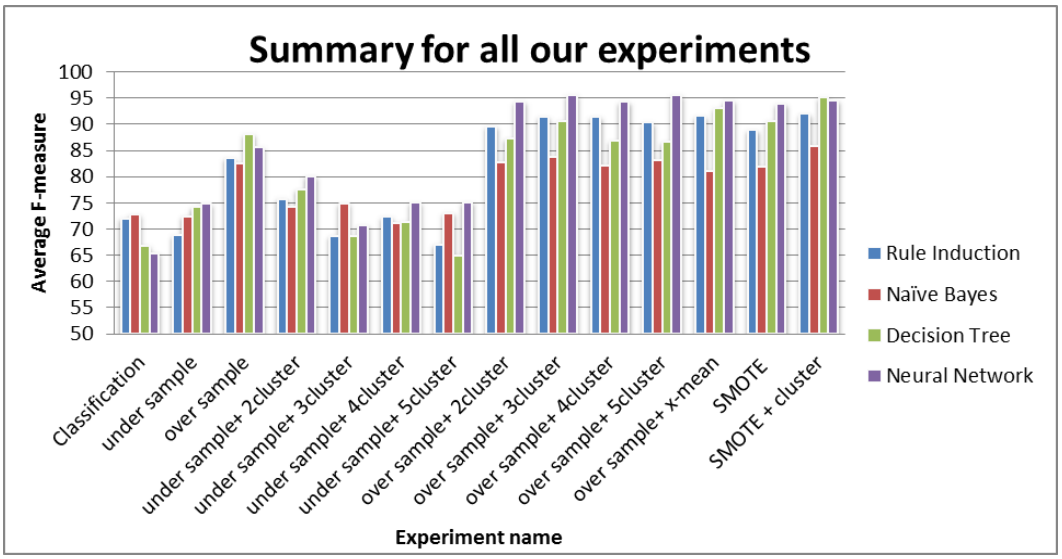

Figure 1: Summary for all our experiments.

We find under sample approach is good solution for imbalanced data distribution but the over sample approach is better than under sample approach because it is difference than under sample approach so there is no information is lost, all instances are employed. Also we preferred use clustering with both two samples approach: under and over sample because we find some of kind of distribution between the data inside the cluster that is helping us in covers all the characteristics of all the existing data. And that when select the majority class samples from each cluster in under sample approach and addingthemto new data set, or when select minority class examples from each cluster in over sample approach and then replicating the selected examples and adding them to new data set.

From all our experiments we can say the over sample with optimal number of clusters achieved the good classification accuracy of minority class. Also, we note using SMOTE based on clustering perform significant improvement on F-measure results and better than over sample based on clustering approach in most cases. Table (2) illustrates accuracy and F-measure results for the baseline and our approach experiments with all classifier.

Table 2: Average accuracy and F-measure comparison of the approaches: baseline and SOMTE with clustering experiments for all our data set

Classifier	Accuracy of baseline	Accuracy ofour approach	F-measure of baseline	F-measure of our approach
Rule Induction	86.75	**92.06**	72.15	**92.07**
Naïve Bayes	82.55	**85.51**	72.82	**86.02**
Decision Tree	84.46	**95.34**	66.91	**95.18**
Neural Network	87.04	**94.68**	65.47	**94.62**

We can note the great difference in improvement before preprocessing and after apply our approach in accuracy and F-measure. For example, in the decision tree the accuracy is 84.46 and the F-measure is 66.91 in the baseline experiment, and after we apply our approach we obtain 95.34 for accuracy and 95.18 for F-measure.

Also we can show the great difference in improvement with different classes in page block data set as in Table (3) and Auto-MPG data set as in Table (4). In page blocks data set all classes except class 1 are consider as a minority class and note that this data set has high imbalances, because class 1 represents 89.8% from all data and other classes represent the remained. In Auto-MPG data setall classes except class 4 are consider as a minority class especially class 3 and 5. Class 3 and 5 has only 4 and 3 instances respectively.

Table 3: F-measure results – page blocks

Classifier	Experiment name	Class 1	Class 2	Class 4	Class 5	Class 3	F-measure
Rule Induction	Baseline	99.39	73.79	71.43	30	50	75.16
	Our approach	**94.28**	**97.82**	**97.77**	**93.13**	**99.86**	**96.55**
Naive Bayes	Baseline	97.50	69.90	90.48	50	70	75.6
	Our approach	**83.51**	**86.58**	**94.85**	**50.77**	**91.47**	**82.69**
Decision Tree	Baseline	99.93	33.01	76.19	0	20	57.28
	Our approach	**94.41**	**97.68**	**97.64**	**96.36**	**100**	**97.23**
Neural Network	Baseline	98.99	73.79	66.67	50	0	59.5
	Our approach	**92.59**	**96.90**	**97.04**	**94.77**	**99.65**	**96.18**

Table 4: F-measure results – auto-mpg

Classifier	Experiment name	Class 8	Class 4	Class 6	Class 3	Class 5	F-measure
Rule Induction	Baseline	100	100	91.67	0	0	58.13
	Our approach	**100**	**89.29**	**96.61**	**100**	**100**	**96.26**
Naive Bayes	Baseline	100	93.65	70.83	50	0	68.49
	Our approach	**100**	**93.55**	**88**	**100**	**100**	**96.5**
Decision Tree	Baseline	100	100	87.50	50	0	72.33
	Our approach	**100**	**96.77**	**98.00**	**100**	**100**	**98.89**
Neural Network	Baseline	100	98.41	100	0	0	58.97
	Our approach	**98.46**	**95.16**	**96.00**	**100**	**100**	**98.88**

From all the above, experimental results confirm our findings which are saying the SMOTE based on clustering achieved best classification accuracy of minority class in imbalanced class distribution problem with both two and multi classes cases. Because the two class problem is a special case from multi class problem.

5. Conclusion and Future Work

Many of real-world applications are encountered the class imbalanced problem. It is occur when there are many more instances of some classes than others. In such cases, standard classifiers tend to be overwhelmed by the large classes and ignore the small ones. Our research proposes a new approach combine between both Synthetic Minority Over-sampling TEchnique (SOMTE) approach and clustering approach which is able to deal with multi class imbalanced data problem. First, we cluster all the training samples in to some clusters. Then we compute the number instances of each class in all clusters. If a cluster has more majority class samples and less minority class samples, it will behave like the majority class samples. After that in each cluster we apply the SMOTE approach which is generate new synthetic minority instances by interpolating between several minority examples that lie close together. Finally, combine between whole classes to produce new balance training data set.

For our experiments, six data sets are chosen from different real domain, characteristics and sizes. Five from data sets (page blocks, cardiotocography, car evaluation, auto MPG and glass identification) represent multi class problem case and the other one data set (breast cancer-w) represent two class problem case. For evaluation purpose, we use cross-validation method provided by RapidMinerenvironment. Also we assume that the ratio of the number of majority class samples to the number of minority class samples in the training data is set to be 1:1. Experimental results show the SMOTE based on clustering approach perform significant improvement on F-measure results and better than normal over sample based on clustering approach in most cases. In some case F-measure improved from 66.91 to 95.18.

In future work, we will need to find solution of the size of data set that will be increasing when adding new instances with amount close to majority size to create balance data set. This is considering problem especially when dealing with very large data sets. Also we can extend our method to deal with within class imbalance problem. Also, we need to consider the problem of imbalance data with noisy dataset especially if the noise in class attribute. Another direction could be working with data types other than numbers and categories such as multimedia data.

References

[1] A. Adam, I. Shapiai, Z. Ibrahim, M. Khalid, L. Chun Chew, L. WenJau and J. Watada; "A Modified Artificial Neural Network Learning Algorithm for Imbalanced Data Set Problem". cicsyn, pp.44-48, 2010 2nd International Conference on Computational Intelligence, Communication Systems and Networks, (2010).

[2] D. Pelleg A. and Moore; "X-means: Extending K-means with Efficient Estimation of the Number of Clusters". ICML (2000).

[3] G. Nguyen, A. Bouzerdoum, and S. Phung; "Learning pattern classification tasks with imbalanced data sets". In P. Yin (Eds.), Pattern recognition (pp. 193-208). Vukovar, Croatia: In-The, (2009).

[4] M. Galar, A. Fernández, E. Barrenechea, H. Bustince and F. Herrera; "A Survey on Ensembles for Class Imbalance Problem: Bagging, Boosting and Hybrid Based Approaches". IEEE Transactions on System, Man and Cybernetics - Part C: Applications and Reviews, doi: 10.1109/TSMCC.2011.2161285, (2012)

[5] M. Wasikowski and X. Chen" Combating the small sample class imbalance problem using feature selection". IEEE Transactions on Knowledge and Data Engineering, vol. 22, no. 10, pp. 13881400, (2010).

[6] N. Chawla, K.Bowyer, L. Hall, and W.P. Kegelmeyer; "SMOTE: synthetic minority over-sampling technique". In International Conference on Knowledge Based Computer Systems, (2002).

[7] S. Chen, G. Guo and L. Chen "A New Over-Sampling Method Based on Cluster Ensembles". AINA Workshops (2010).

[8] S.-J. Yen and Y.-S.Lee; "Cluster-based Under-sampling Approaches for Imbalanced Data Distributions". Expert Systems with Applications, 36, 5718-5727,(2009).

[9] T. Debray; "Classification of Imbalanced Data Sets".Master's Thesis in Artificial Intelligence Faculty of Humanities and Sciences, Maastricht University, (2009)

[10] V. García, J.S. Sánchez, R.A. Mollineda, R. Alejo, and J.M. Sotoca" The class imbalance problem in pattern classification and learning". Tamida, Saragossa, Spain, pp. 283-291, (2007).

[11] Y. Murphey, H. Wang, G. Ou and L. Feldkamp; "OAHO: an effective algorithm for multi-class learning from imbalanced data". in International Joint Conference on Neural Networks (IJCNN), pp. 406–411, (2007).

[12] Y. Sun; "*Cost-Sensitive Boosting for Classification of Imbalanced Data*". Thesis requirement for the degree of Doctor of Philosophy In Electrical and Computer Engineering, Waterloo University, (2007).

[13] UCI Machine Learning Repository: Page Blocks Data set, Available: http://archive.ics.uci.edu/ml/datasets/Page+Blocks+Classification, (2012, March), [Online].

[14] UCI Machine Learning Repository: Cardiotocography Data set, Available: http://archive.ics.uci.edu/ml/datasets/Cardiotocography, (2012, March), [Online].

[15] UCI Machine Learning Repository: Car Evaluation Data set, Available: http://archive.ics.uci.edu/ml/datasets/Car+Evaluation, (2012, March), [Online].

[16] UCI Machine Learning Repository: Auto MPG Data set, Available: http://archive.ics.uci.edu/ml/datasets/Auto+MPG, (2012, March), [Online].

[17] UCI Machine Learning Repository: Glass Identification Data set, Available: http://archive.ics.uci.edu/ml/datasets/Glass+Identification, (2012, March), [Online].

[18] UCI Machine Learning Repository: Breast Cancer Wisconsin Data set, Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Original%29, (2012, March), [Online].