Improving Imbalanced Students' Text Feedback Classification Using Re-sampling Based Approach

Zainab.Mutlaq Ibrahim, Mohamed Bader-El-Den, Mihaela Cocea
(zainab.mutlaq-ibrahim, mohamed.bader, mihaela.cocea)@port.ac.uk

University of Portsmouth, Lion Terrace PO1 3HE, UK

Abstract: **Class imbalance is a major problem in text classification, the problem happens when the used machine learning algorithm biases towards the majority class, so this makes it incorrectly classifies minority class instances. To get over this problem investigators use the Synthetic Minority Oversampling Technique(SMOTE), it is preprocessing algorithm which was proven as a very good solution for handling imbalanced data sets. In this paper an empirical study have been executed to handle three imbalanced data sets in text format using SMOTE, the recall of all minority classes significantly improved in addition of significant improvement in all models overall performance.**
Average classes' recall was improved significantly, by 0.15, 0.09, 0.10 in classification of ASS, FDS, NASS data sets respectively. While the recall for the minority class has significantly increased, ASS(0.23), FDS(0.08, and NASS(0.15)
keywords: imbalanced data set, SMOTE, text feedback, text classification

1 Introduction

Class imbalance is considered to be a big problem in text classification, the problem appears when classes do not make up an equal portion of a data-set. For example, in a simple two classes case, a balanced state would have the class priority of both classes approximately equal to each other.
In case of an imbalanced problem, the majority class has much more priority than the minority class.
It is very important to correctly classify all instances in a data set whether they belong to majority or minority class, in some cases, it is costly to dis-classify minority instances such as in cancerous cells[1], breast and colon cancer[2], this miss-classification can lead to wrongly diagnosis the illnesses, mess up and delay efficient and quick treatment in both cases.
Wrong classification in Fraud detection[3,4], information retrieval[5], marketing [6], and keyword extraction [7] can lead to frauds success, wrong decision and results or marketing the wrong product. In addition, detecting oil-spill wrongly can lead to environment pollution and harming the nature [8].

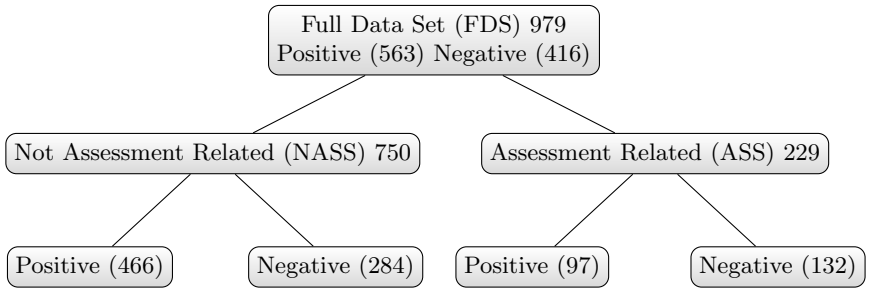
Without considering imbalanced priorities, a classifier may learn to always predict the majority class. The cost of incorrectly classifying the minority class may extremely high and not acceptable[9,10].
Approaches have been proposed to deal with class imbalance problem include re-sampling techniques which changes the priors in the training set by either generating more instances of minority class or omitting instances from majority class.
More techniques for dealing with class imbalanced include appropriate feature selection [11], cost-sensitive learners that consider miss-classification cost in the learning phase[12], one class learner[13,10], and hybrid of the above techniques.
In this paper we are going to use re-sampling method because it showed a good results [14], researches have shown a strong relation between re-sampling methods and cost-sensitive cost[12], and also re-sampling method is easy to implement.
The rest of the paper includes:Section 2 which outlines some related work that deal with class imbalance, section 3 describes and illustrates the framework, section 4 details the empirical study, and section 5 concludes the paper.

2 Related Work

Ferdenands and et al explained different techniques to deal with class imbalance problem, two of them refer to oversampling and down sampling the data set [15]. oversampling methods have been proposed and adopted in different studies, Synthetic Minority Oversampling Technique(SMOTE) is an oversampling technique, in this technique more instances of minority class are generated.
Mohasseb and et al improved Naive bays classifier performance significantly[16] by implementing such a method, while Sisovic and et al slightly improve their clustering model performance [17]. Awad and et al enhanced the performance of all models using SMOTE method[18]. LV and et al used SMOTE to improve the recognition of their model which is to distinguish and recognize users who steal electricity[19].
Down sampling technique is to take out a set of the the majority class to balance the data set, this technique is used more often in image processing, Wang and ec al used this technique to get image super-resolution [20], and Lin and ec al used it to improve image compression at low bit rates [21].

3 Data Distribution

The used data set in this study was collected from school of computing/university of Portsmouth between 2012-2016 as end of unit feedback, for more details about the data set please see [22].
Figure. 1. shows the imbalanced class distribution(positive, negative), in the full data set(FDS), assessment related(ASS), and not assessment related (NASS)



.

Fig. 1. Data Set Distribution[22]

4 Proposed framework

Figure 2 illustrates the proposed framework of dealing with imbalanced data set using SMOTE method.
The proposed framework first transforms the text documents into structured data, there is an imbalanced problem regarding sentiment classes (positive, negative) as illustrated in figure 1 section 2, so the framework has two routes, first follows the 'yes' route after binary question whether the data set is balanced or not (the black arrows) to conclude result 1, and second follows the gray arrows to conclude result 2 .
Text pre-processing can include remove numbers, punctuation marks, stop words and words that less than three letters in addition to apply the SMOTE algorithm. SMOTE method works better with binary labels[16], this make it perfect for our data sets, it runs an oversampling approach to re-balance the original training set. it applies a simple reproduction of the minority class instances. The main task of SMOTE is to introduce artificial and unreal examples. This new data is created by insertion between several minority class instances that are within a defined neighborhood.
In this paper, Support vector machine algorithm was used to build the classification model, it is a powerful tool for text classification, it has the power to determine an optimal separating point that labels records into different categories.
The final Phase is to evaluate models in both routes to see and compare their performance using recall precision, accuracy, and F-Measure.

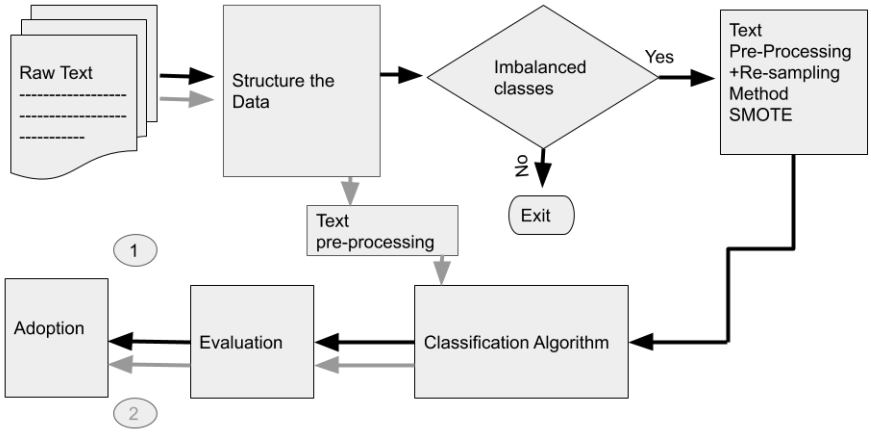


Fig. 2. Proposed Framework for Handling imbalanced data set

5 Empirical Study

In all experiments to build the needed models, a computer desktop was used with quad 2.33 GHZ CPU, 4GB RAM, and windows 7 operating system.
WEKA a graphic user interface(GUI) was used to pre-proccess, classify, and re-sample our data sets.
Support Vector Machine(SVM) was used as a machine learning algorithm for automatic text classification, it was applied using data set and its subsets that used in[22] to build three models, the performance of these models evaluated using Accuracy ,Precision, Recall, and F-Measure measurements, the study used 10-fold cross validation.

This study results show the success of handling imbalanced data of classification models' performance using SMOTE technique, first this paper shows the results of applying SVM algorithm in classification process without using SMOTE, results were listed in [22] and second rerun the experiment using smote technique, see table 1

5.1 Results

Table 1 presents overall classification performance details of SVM classifier using the SMOTE algorithm and without using it.The results show better classifier performance when handling the class imbalance.
To shed the light in how did the classifier classify the minority class in the above data sets we need to have a look at the recall results which are illustrated in

	SVM without SMOTE				SVM with (SMOTE)			
Data set	**Accuracy**	**Precision**	**Recall**	**F-Measure**	**Accuracy**	**Precision**	**Recall**	**F-Measure**
ASS	0.69	0.70	**0.70**	0.69	0.85	0.88	**0.85**	0.84
FDS	0.76	0.76	**0.74**	0.74	0.83	0.83	**0.83**	0.83
NASS	0.76	0.75	**0.73**	0.74	0.85	0.88	**0.85**	0.84

Table 1. Overall SVM classifier performance without/with the implementation of SMOTE algorithm

table 2. The results shows that classification of the minority class in all data set has improved, in FDS data set, the recall increased by 0.08, in the NASS data set case the recall increased by 0.15 while in ASS data set the recall significantly increased by 0.23 in total.

	SVM without SMOTE			SVM with (SMOTE)		
	FDS	NASS	ASS	FDS	NASS	ASS
P/N	563/416	466/284	97/132	563/416	466/284	97/132
P	0.85	0.84	**0.64**	0.92	0.75	**0.87**
N	**0.63**	**0.70**	1	**0.71**	**0.85**	0.79

Table 2. Recall SVM classifier performance without/with the implementation of SMOTE algorithm P=Positive,N=negative

6 Conclusion and Future Work

This study proposed a framework for handling class imbalance issue using SMOTE algorithm and utilizing Uni gram feature.
Empirically, results have shown that the proposed framework worked well for the used data set classification, and improved all models in terms of accuracy, precision, recall, and F-measure.
Average classes' recall was improved significantly, by 0.15, 0.09, 0.10 in classification of ASS, FDS, NASS data sets respectively.
The recall for the minority class has significantly increased using SMOTE for all data sets, the best recall improvement was in ASS data set(positive class), it increased by 0.23, the second was in NASS(negative class) data set which increased by 0.15, and finally FDS (negative class) by 0.08.
Future work include apply more techniques for dealing with class imbalanced such as appropriate feature selection, cost-sensitive learners that consider mis classification cost in the learning, and the SMOTE algorithm with different feature representations such as bi-gram and Part of Speech(PoS).

References

1. Rushi Longadge and Snehalata Dongre. Class imbalance problem in data mining review. *arXiv preprint arXiv:1305.1707*, 2013.
2. Abdul Majid, Safdar Ali, Mubashar Iqbal, and Nabeela Kausar. Prediction of human breast and colon cancers from imbalanced data using nearest neighbor and support vector machines. *Computer methods and programs in biomedicine*, 113(3):792–808, 2014.
3. Clifton Phua, Damminda Alahakoon, and Vincent Lee. Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter*, 6(1):50–59, 2004.
4. Philip K Chan and Salvatore J Stolfo. Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection. In *KDD*, volume 1998, pages 164–168, 1998.
5. Peter D Turney. Learning algorithms for keyphrase extraction. *Information retrieval*, 2(4):303–336, 2000.
6. Charles X Ling and Chenghui Li. Data mining for direct marketing: Problems and solutions. In *Kdd*, volume 98, pages 73–79, 1998.
7. David D Lewis and William A Gale. A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer, 1994.
8. Miroslav Kubat, Robert C Holte, and Stan Matwin. Machine learning for the detection of oil spills in satellite radar images. *Machine learning*, 30(2-3):195–215, 1998.
9. Alexander Liu, Joydeep Ghosh, and Cheryl E Martin. Generative oversampling for mining imbalanced datasets. In *DMIN*, pages 66–72, 2007.
10. Shiven Sharma, Colin Bellinger, Bartosz Krawczyk, Osmar Zaiane, and Nathalie Japkowicz. Synthetic oversampling with the majority class: A new perspective on handling extreme imbalance. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 447–456. IEEE, 2018.
11. Zhaohui Zheng, Xiaoyun Wu, and Rohini Srihari. Feature selection for text categorization on imbalanced data. *ACM Sigkdd Explorations Newsletter*, 6(1):80–89, 2004.
12. Bianca Zadrozny, John Langford, and Naoki Abe. Cost-sensitive learning by cost-proportionate example weighting. In *ICDM*, volume 3, page 435, 2003.
13. Bhavani Raskutti and Adam Kowalczyk. Extreme re-balancing for svms: a case study. *ACM Sigkdd Explorations Newsletter*, 6(1):60–69, 2004.
14. Guillem Collell, Drazen Prelec, and Kaustubh R Patil. A simple plug-in bagging ensemble based on threshold-moving for classifying binary and multiclass imbalanced data. *Neurocomputing*, 275:330–340, 2018.
15. Alberto Fernández, Salvador García, Mikel Galar, Ronaldo C Prati, Bartosz Krawczyk, and Francisco Herrera. *Learning from imbalanced data sets*. Springer, 2018.
16. Alaa Mohasseb, Mohamed Bader-El-Den, Mihaela Cocea, and Han Liu. Improving imbalanced question classification using structured smote based approach. In *2018 International Conference on Machine Learning and Cybernetics (ICMLC)*, volume 2, pages 593–597. IEEE, 2018.
17. Sabina Šišović, Maja Matetic, and Marija Brkic Bakaric. Clustering of imbalanced moodle data for early alert of student failure. pages 165–170, 01 2016.
18. Aya Awad, Mohamed Bader-El-Den, James McNicholas, and Jim Briggs. Early hospital mortality prediction of intensive care unit patients using an ensemble

learning approach. *International Journal of Medical Informatics*, 108:185 – 195, 2017.

19. Dong Lv, ZhiCheng Ma, Shibo Yang, Xianbo Li, Zhixin Ma, and Fan Jiang. The application of smote algorithm for unbalanced data. In *Proceedings of the 2018 International Conference on Artificial Intelligence and Virtual Reality*, pages 10–13. ACM, 2018.
20. Yifan Wang, Lijun Wang, Hongyu Wang, and Peihua Li. Information-compensated downsampling for image super-resolution. *IEEE Signal Processing Letters*, 25(5):685–689, 2018.
21. Weisi Lin and Li Dong. Adaptive downsampling to improve image compression at low bit rates. *IEEE Transactions on Image Processing*, 15(9):2513–2521, 2006.
22. Zainab Mutlaq Ibrahim, Mohamed Bader-El-Den, and Mihaela Cocea. Mining unit feedback to explore students' learning experiences. In *UK Workshop on Computational Intelligence*, pages 339–350. Springer, 2018.