International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:6, 2016

An Analysis of Classification of Imbalanced Datasets by Using Synthetic Minority Over-Sampling Technique

Ghada A. Alfattni

***Abstract*—**Analysing unbalanced datasets is one of the challenges that practitioners in machine learning field face. However, many researches have been carried out to determine the effectiveness of the use of the synthetic minority over-sampling technique (SMOTE) to address this issue. The aim of this study was therefore to compare the effectiveness of the SMOTE over different models on unbalanced datasets. Three classification models (Logistic Regression, Support Vector Machine and Nearest Neighbour) were tested with multiple datasets, then the same datasets were over-sampled by using SMOTE and applied again to the three models to compare the differences in the performances. Results of experiments show that the highest number of nearest neighbours gives lower values of error rates.

***Keywords*—**Imbalanced datasets, SMOTE, machine learning, logistic regression, support vector machine, nearest neighbour.

I. Introduction

IMBALANCED dataset is a term used to describe data when its classes are not approximately equal. Dealing with this issue is one of the challenges that practitioners in machine learning field face. However, two ways have been used to address this problem. One way is to assign distinct costs to training examples (this solution has not been discussed in this paper due to time constraints), while the second one is re-sampling a dataset by under-sampling the majority class or over-sampling the minority class [1]. This project will examine SMOTE, which is used to over-sample the minority class to address the imbalanced datasets issue.

This paper is divided in the following sections: Methods section briefly describes SMOTE algorithm and its common properties in addition to formal techniques that have been used during the investigation. Experiments section reports the experimental work. Analysis section analyses the experimental results. The last two sections provide some conclusions and future work.

II. Methods

We examine the three machine learning classifiers before and after re-sampling the datasets by SMOTE.

A. Classifiers

For each classifier, the model starts by randomising the data so that the results are independent. Then, cross validation has been done by randomly dividing the datasets into five parts; four parts for training and one for the testing. Afterwards, error rate and standard deviations have been calculated in order to use them as comparison criteria as suggested by [2].

B. SMOTE

SMOTE is a technique used to deal with the problem of having few cases in the minority class in a dataset [3]. The module creates new examples by taking minority class samples and "introduces synthetic examples along the line segments joining any/all of the k minority class nearest neighbours" [1]. Our investigation tests the model before applying SMOTE to the datasets and after applying SMOTE with 3 and 5 nearest neighbours. Fig. 1 shows an example of using SVM classifier with and without applying the SMOTE to a dataset.

C. Evaluation Criteria

As previously mentioned, error rates and standard deviations were considered as evaluation criteria in this paper. Error rate is needed to denote the inaccuracy of predicted output values [4]. Standard deviation is needed as well in order to quantify the dispersion of the data from its mean [5].

Although, error rate is one of the criteria that measures the performance and could affect our decision, we also introduce the confusion matrices for all models to get the accuracy, precision, recall and f-measure for each classifier. Accuracy is needed to give us the overall view on how often the classifier is correct, while f-measure gives an average of the true positive rate (recall) and precision [6].

The last and most important measurement criterion that was made is the Receiver Operating Characteristic (ROC) curve, which plots the true positive rate against the false positive rates. It was used in our comparison to provide an overview of the optimal model.

III. Experiments

A. Datasets

Three datasets taken from MLOtools were used in this experiment with slightly changes. The first one is Muddleddata, which consists of 100 examples and 2 features. The second one is Heartt dataset, and it consists of 270 examples and 13 features. Congress is the last dataset used and it consists of 435 examples and 16 features. Each example in the three datasets is labelled with one of two classes. Table I describes the characteristics of each dataset.

Ghada A. Alfattni is with the Um Alqura University, Makkah, Saudi Arabia (e-mail: gafattni@uqu.edu.sa).

B. Parameters Tuning

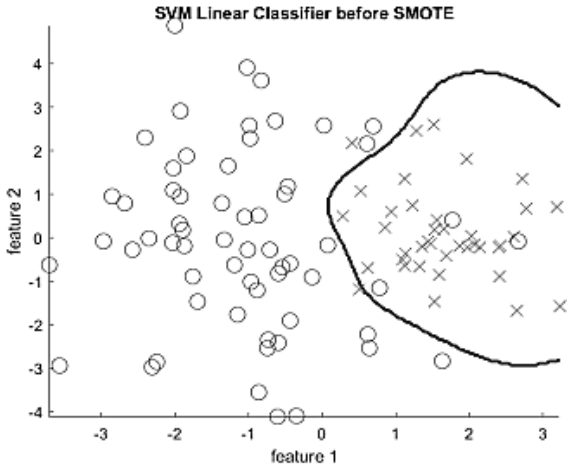
Logistic Regression and KNN classifiers have implemented with the default parameters. Learning rate = 0.1 and iterations = (10 x number of features) for Logistic Regression, and Euclidean distance k = 3 for KNN. However, SVM has implemented with slack variable C = 2 and a radial basis function (1) for all datasets.

$$\exp(-\gamma \|u - v\|^2) \quad (1)$$

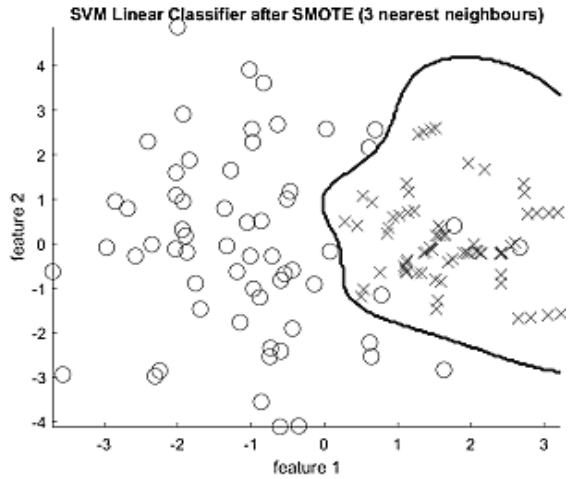
SMOTE algorithm was used two times in this analysis. First, it was used to find 3 nearest neighbours of all points in the minority classes, and then it was used to find 5 nearest neighbours.

TABLE I
CHARACTERISTICS OF DATASETS

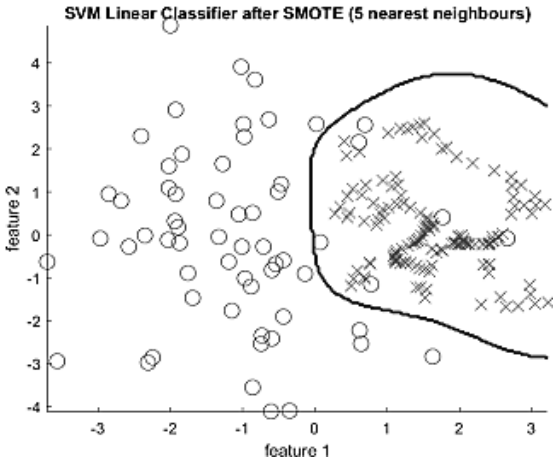
Dataset	Examples	Features	Classes	Majority class	Minority class
Muddleddata	100	2	2	61	39
Heartt	270	13	2	150	120
Congress	435	16	2	267	168



(a)



(b)



(c)

Fig. 1 Example of using SVM classifier before and after applying the SMOTE to a dataset; (a) Under SMOTE; (b) SMOTE (3 nearest neighbours); (c) SMOTE (5 nearest neighbours)

C. Performance Evaluation

To estimate the generalization performance, 5-folds cross-validation was performed on random permutations of the datasets for each classifier. Means and standard deviation were calculated and used to plot bar charts containing error bars and standard deviations. Results are shown in Table II and Fig. 2. Confusion matrices and its common metrics were calculated as shown in Table III and Fig. 3.

TABLE II
MEAN OF GENERALIZATIONS ERROR FOR THE DATASETS

Logistic Regression			
Dataset	Under SMOTE	SMOTE (3)	SMOTE (5)
Muddleddata	0.1000	0.0595	**0.0378**
Heartt	0.1815	0.1222	**0.0822**
Congress	0.0644	0.0210	**0.0033**
SVM			
Dataset	Under SMOTE	SMOTE (3)	SMOTE (5)
Muddleddata	0.1100	0.0431	**0.0279**
Heartt	0.1704	0.0982	**0.0785**
Congress	0.0552	0.0123	**0.0044**
KNN			
Dataset	Under SMOTE	SMOTE (3)	SMOTE (5)
Muddleddata	0.1000	0.0364	**0.0340**
Heartt	0.2074	0.0445	**0.0385**
Congress	0.0782	0.0298	**0.0285**

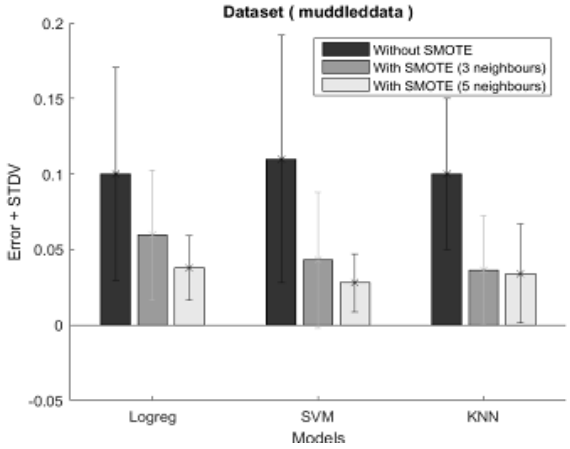
IV. ANALYSES

As it is seen in Fig. 2, we can notice that the classifiers' behaviors are the same on all datasets. SMOTE provides a better estimation with the imbalanced datasets and gives a minimum error rate even if the differences between the majority and minority classes are not fixed. In the same figure, if we look at the standard deviations in the three datasets, we can notice that, when the number of nearest neighbours increased, the differences between the standard deviations in each classifier are statistically significant. To clarify, the

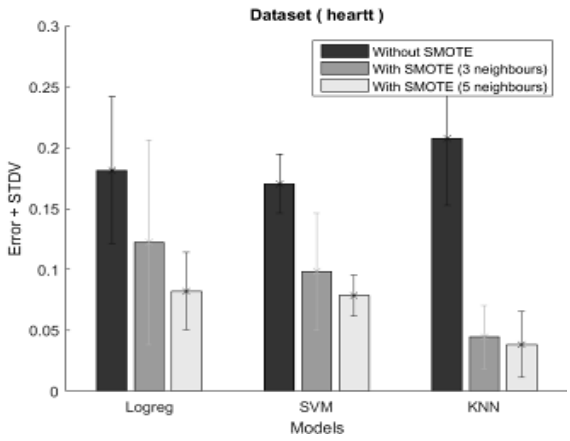
values of standard deviations are decreased to almost the half when the SMOTE is not used, which is surprising. Moreover, this tells us that the classification after applying SMOTE technique to the dataset will give a good model.

Although the standard deviations of the error rates are decreased after using SMOTE, they are still at high values. This is because the datasets that have been used in this study were relatively small (100, 270, and 435), which might increase the noise in result of the classifier [7].

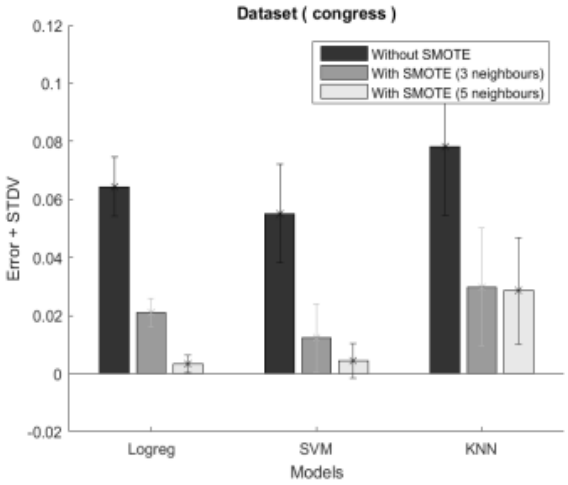
Moving to the results shown in Table III, we can see that all classifiers make the highest amount of correct predictions after SMOTE. While SVM model gave the same value of accuracy for SMOTE with 3 and 5 nearest neighbours. The values of F-measure are also noticeable here. Generally, it reaches the highest value after SMOTE. That is to say, the average of the true positive rate (recall) and precision reaches its best value when we apply SMOTE technique to the dataset before training it and testing it by the classifier.



(a)



(b)



(c)

Fig. 2 Bar plot of the error rates and standard deviations; (a) Muddleddata dataset; (b) Heartt dataset; (c) Congress dataset

TABLE III
ACCURACY AND F-MEASURE FOR EACH CLASSIFIER IN THE THREE DATASETS

Muddleddata Datasets				
Classifier		Under SMOTE	SMOTE (3)	SMOTE (5)
Logistic Regression	Accuracy	0.9000	0.9403	**0.9624**
	F-measure	0.9180	**0.9322**	0.9298
SVM	Accuracy	0.8900	**0.9552**	**0.9552**
	F-measure	0.9091	**0.9493**	0.9483
KNN	Accuracy	0.9000	0.9627	**0.9718**
	F-measure	0.9167	**0.9573**	0.9492
Heartt Datasets				
Classifier		Under SMOTE	SMOTE (3)	SMOTE (5)
Logistic Regression	Accuracy	0.8185	0.8783	**0.9181**
	F-measure	0.7841	0.9198	**0.9492**
SVM	Accuracy	0.8296	**0.9021**	**0.9021**
	F-measure	0.8034	**0.9296**	**0.9296**
KNN	Accuracy	0.7926	0.9555	**0.9808**
	F-measure	0.7647	0.9684	**0.9886**
Congress Datasets				
Classifier		Under SMOTE	SMOTE (3)	SMOTE (5)
Logistic Regression	Accuracy	0.9356	0.9791	**0.9967**
	F-measure	0.9205	0.9821	**0.9978**
SVM	Accuracy	0.9448	**0.9878**	**0.9878**
	F-measure	0.9310	0.9896	**0.9896**
KNN	Accuracy	0.9218	0.9703	**0.9867**
	F-measure	0.9017	0.9750	**0.9912**

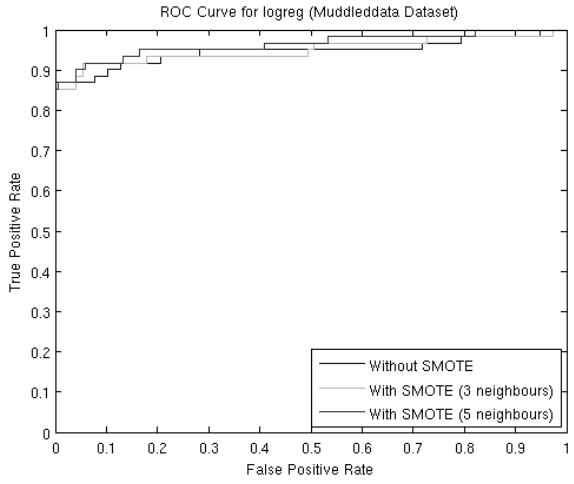
The final performance criterion that we would like to analyze in this study is the ROC curve. We know that the best possible prediction model would create a point in the upper left corner the ROC space. Fig. 3 illustrates the ROC curves for each dataset applied to the logistic regression model before SMOTE, after SMOTE with 3 nearest neighbours and SMOTE with 5 nearest neighbours. By looking at it, we can conclude that the SMOTE generally gives a good separation between classes. In Heartt dataset, despite the fact that the ROC curve for the data before SMOTE is mostly convex, it

becomes more separable after SMOTE.

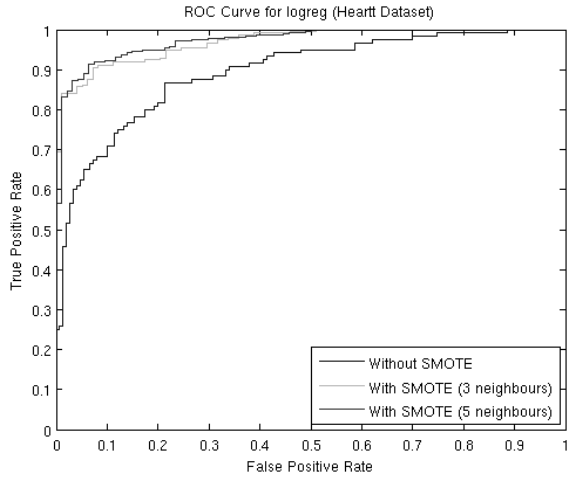
V. Conclusions

The aim of this project was to evaluate the SMOTE technique and investigate its performance on the given datasets. Experiments focused on the three machine learning algorithms: Logistic Regression, Support Vector Machine and Nearest Neighbour. Increasing the number of nearest neighbours inside the SMOTE to introduce new data points is shown on all datasets. Error rates, standard deviations, accuracy, and F-measure have been calculated to measure the performance of each classifier.

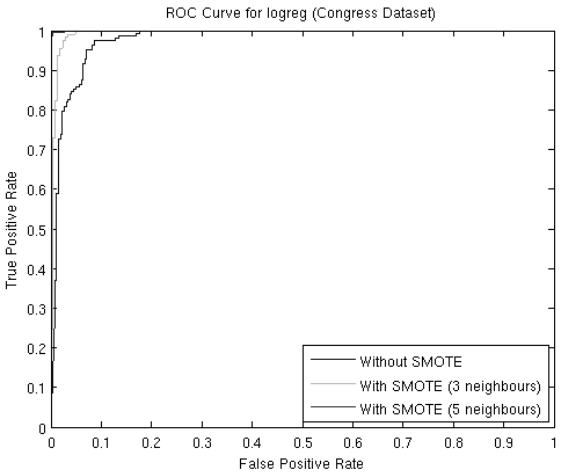
The highest number of nearest neighbours gives lower value of error rates which are shown on error bar charts, besides giving a good separation between classes, which are illustrated in ROC curves. To conclude, over-sampling by SMOTE is suited for the imbalanced datasets and it could improve the classifier's accuracy.



(a)



(b)



(c)

Fig. 3 ROC curves for logistic regression model; (a) Muddleddata dataset; (b) Heartt dataset; (c) Congress dataset

VI. Future Works

There are several research topics that could be considered as a future work. Comparing between assigning distinct costs to train examples and re-sampling a dataset by under-sample the majority class or over-sample the minority class as they are two solutions to deal with imbalanced datasets. It is also possible to examine relatively bigger amount of imbalanced datasets.

References

[1] Chawla, N.V., et al., *SMOTE: Synthetic Minority Over-Sampling Technique.* Journal of Artificial Intelligence Research, 2002. 16: p. 321-357.

[2] Alpaydin, E., *Introduction to Machine Learning*. 2009: Massachusetts Institute of Technology.

[3] Dong, Y. and X. Wang, *A New Over-Sampling Approach: Random-SMOTE for Learning from Imbalanced Data Sets.* KSEM'11 Proceedings of the 5th international conference on Knowledge Science, Engineering and Management, 2011: p. 343-352.

[4] Kohavi, R. and F. Provost. *Glossary of Terms: Special Issue on Applications of Machine Learning and the Knowledge Discovery Process*. 1998 (cited 2016); Available from: http://robotics.stanford.edu/~ronnyk/glossary.html.

[5] Bland, J.M. and D.G. Altman, *Measurement error.* British Medical Journal, 1996. 313: p. 744.

[6] Fawcett, T., *An introduction to ROC analysis.* Pattern Recognition Letters - Special issue: ROC analysis in pattern recognition, 2006. 27(8): p. 861-874.

[7] Brain, D. and G.I. Webb, *On the effect of data set size on bias and variance in classification learning.* The Fourth Australian Knowledge Acquisition Workshop, 1999: p. 117-128.