Modeling of class imbalance using an empirical approach with spambase dataset and random forest classification

By: Kiranmayi Kotipalli and Shan Suthaharan

K. Kotipalli, and S. Suthaharan. 2014. Modeling of class imbalance using an empirical approach with spambase dataset and random forest classication, ACM RIIT 2014, pp. 75-80.

Abstract:

Classification of imbalanced data is an important research problem as most of the data encountered in real world systems is imbalanced. Recently a representation learning technique called Synthetic Minority Over-sampling Technique (SMOTE) has been proposed to handle imbalanced data problem. Random Forest (RF) algorithm with SMOTE has been previously used to improve classification performance in minority class over majority class. Although RF with SMOTE demonstrates improved classification performance, the relationship between the classification performance and the imbalanced ratio between the majority and minority classes is not well defined. Therefore mathematical models that describe this relationship is useful especially in the big data environment which suffers from imbalanced data. In this paper, we proposed a mathematical model using an empirical approach applied to the well known Spambase dataset and Random Forest classification approach including its adoption with SMOTE representation learning technique. We have presented a linear model which describes the relationship between true positive classification rate and the imbalanced ratio between the majority and minority classes. This model can help IT researchers to develop better spam filter algorithms.

Keywords: Random forest | SMOTE | Imbalanced data | Classification | Machine learning

Article:

1. INTRODUCTION

Many real world applications including network intrusion detection, document classification, Spam filtering, fraud detection and drug discovery, suffer from imbalanced data problems consistently.

In these applications, the class that is of interest is under represented, and thus the accurate classification of the minority class than the majority class becomes difficult. For instance, in intrusion detection systems, attack patterns or malicious activities can be classified by monitoring the network where the number of instances of attacks is comparatively much smaller than the regular network traffic. It is therefore extremely challenging to classify such imbalanced data with machine learning techniques that in general learn from the characteristics of the majority class.

Figure 1 shows a representation of imbalanced data using two classes plotted against two attributes (x-axis and y-axis). The minority class is denoted by circles and the majority class denoted by crosses. In this case, the data points of the minority class may be treated as outliers and anomaly detection algorithms may be applied. However, the classification algorithms require balance between the classes and hence it is challenging to derive optimal classifiers when the dataset is imbalanced.

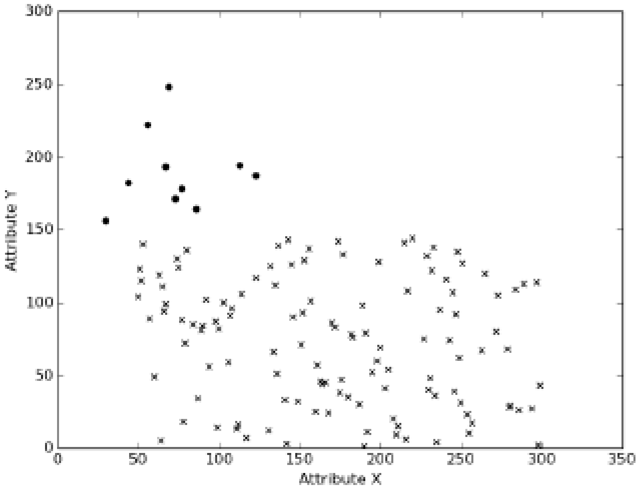


Figure 1. Imbalanced data with two classes

There are two forms of class imbalance problems in machine learning areas [8]: between-class imbalance data (a commonly occurring problem where the majority class samples out represent the minority class), and within-class imbalance data (it occurs when there exist small clusters of data within a class that are under represented).

When classification algorithm such as C4.5 or any learner in general is applied on imbalanced data, it is more likely to classify the minority class as the majority class [13]. Thus machine learning algorithms like Random Forest (RF) [10], Support Vector Machine (SVM) [1], Deep Learning (DL) [9] may also be biased to majority class. This is because these algorithms are first trained on the class data which has fewer samples of minority data and therefore will be more biased to the majority class. Also, since the data is divided into training and test samples, the probability of bias to the majority class can be even higher. Foster Provost [11] attributed this problem to the assumptions made by the machine learning algorithms.

To handle the imbalanced data problem in classification, a representation learning technique called Synthetic Minority Oversampling Technique (SMOTE) [5], which can be adopted with a classification technique like RF, has recently been proposed in machine learning. The effectiveness of RF combined with SMOTE has been reported in machine learning, but modeling of class imbalance and its effect on the performance of RF or RF with SMOTE is still required. Such models will be useful to address big data problems and develop automated tools for information technology.

In this paper we conducted an empirical research using RF, RF with SMOTE and Spambase dataset (available at UCI repository) [2], and modeled the relationship between the variability in minority and majority classes, and the true positive classification rate. The model derived is linear and hence it is useful for automating big data classifiers to handle imbalanced data problems.

2. RANDOM FOREST AND SMOTE

Random Forest is a machine learning algorithm that uses an ensemble approach by combining many decision tree models. To grow these decision trees, firstly different subsets of data are randomly generated from the original dataset with replacement. This is called bootstrap aggregation or bagging. These subsets of data are then fed to individual decision trees that classify the data by selecting a random subset of the features at each split. The best split at each node is selected using the GINI impurity. Each tree then casts its vote on a class. Random forest then uses the majority vote among all trees to classify the data. In this way each individual tree acts as a weak classifier and combines with all other trees in the forest to become a strong classifier [3]. When a subset of data is used to train a decision tree, the remaining data which is called the out-of-bag sample is used to estimate error and variable importance [4].

Many techniques and algorithms have been proposed to improve the original Random Forest algorithm on imbalanced data. Some of these methods include modifying the imbalanced dataset to balance the data (sampling technique) and associating high cost for misclassification of minority class (cost-sensitive learning) [6]. Oversampling and undersampling are some of the commonly used sampling techniques. Oversampling involves duplicating randomly selected minority class samples, while undersampling involves selecting a small random subset of the majority class for training. Both these techniques balance the data and are simple to implement. However oversampling introduces the problem of overfitting and undersampling results in loss of information of the majority class. Many of these algorithms have been tested on imbalanced data where the class distribution of minority class may range from 1% to 50%. There is no benchmark on what is the percentage of class distribution that really makes a class imbalanced for classification, and this is the focus of our research presented in this paper.

SMOTE is an oversampling approach in which the minority class is oversampled by creating synthetic or artificial samples instead of oversampling with replacement. It is based on the idea that the samples closer to the minority class also belong to the minority class. This is achieved by introducing new samples along the line segment joining the k-nearest neighbor minority class which are selected based on the Euclidean distances. Based on the amount of oversampling the nearest neighbors are chosen randomly. After choosing the nearest neighbor, the difference between its feature vector with the current sample is computed and multiplied with a random number between 0 and 1. This value is then added to the feature vector space, thus creating a new feature. This way a new sample is created along the line segment between two specific features.

The default implementation uses five nearest neighbors. So, in order to achieve 100% oversampling, one neighbor among the five nearest neighbors is chosen randomly and a new sample is generated in that direction. Using SMOTE, the decision region of the minority class becomes less specific as it is increased by encompassing the nearest neighbors. This is a better approach than the oversampling with replacement technique because mere data replication creates specific decision regions leading to over fitting problem.

3. SPAMBASE DATA SET

For this experiment the spambase dataset from the UCI repository is considered which was donated by George Forman from Hewlett-Packard laboratories, Palo Alto, California [2]. This dataset contains a collection of mails containing regular and spam mails. Spam mails include unsolicited commercial mail with advertisements, schemes for making money, chain letters etc. Table 1 provides a summary of this dataset.

Table 1. Spambase dataset summary

Number of classes	2
Number of Instances	4601
Number of Spammails(Class1)	1813
Number of Non-spammails(Class0)	2788
Number of Attributes	57

The dataset was created to build a spam filter to distinguish between regular and spam mail. The data for this dataset is collected by the postmaster and individuals _ling spam mail. Most of the attributes indicate whether a particular word or character was frequently occurring in email.

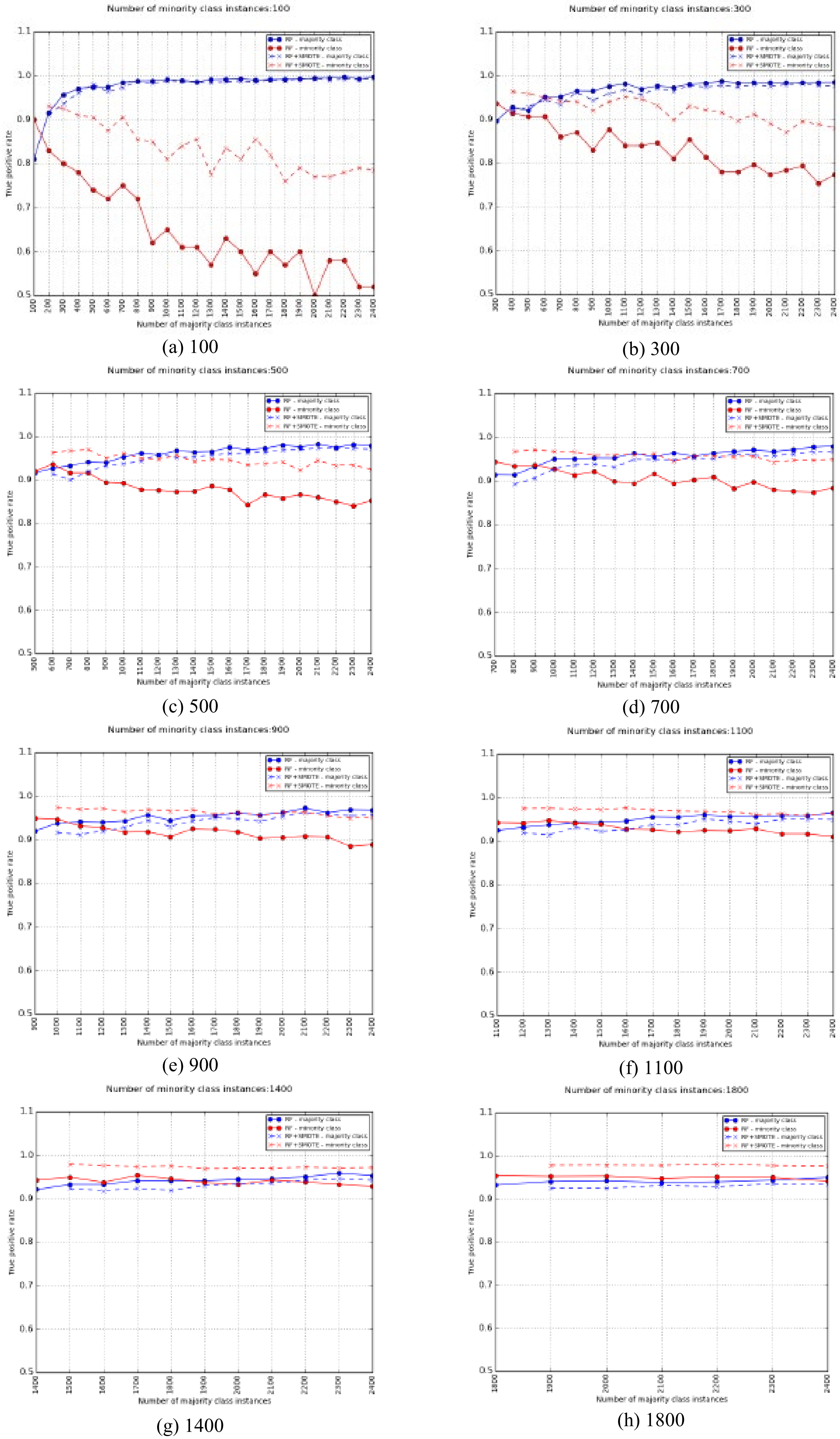
For example, word_freq_money indicates the number of times the word money occurs in a mail. This is given as a percentage of words in the e-mail that match the word money. Table 2 shows the percentages for some of the words. Occurrence of words like George, hp (company name) and 650 (area code) indicate genuine mails while words like free, money and the character ! indicate spam mail. Some of the attributes look for uninterrupted characters.

Table 2. Statistics of words in spambase

	free	!	money	george	hp	650
spam	0.52	0.51	0.21	0	0.02	0.02
legitimate	0.07	0.11	0.017	1.27	0.9	0.19

Similarly, capital_run_length_longest is the length of longest uninterrupted sequences of capital letters. The last attribute type indicates the class: Class 0 indicates legitimate or regular mail and class 1 indicates spam mail.

This dataset is an example of imbalanced data as the ratio of spam to legitimate email is approximately 0.65. The minority class is the spam email and the majority class is the legitimate email.



(a) 100

(b) 300

(c) 500

(d) 700

(e) 900

(f) 1100

(g) 1400

(h) 1800

Figure 2. True positive curve for spambase dataset with different degrees of imbalance

4. SETTINGS FOR EXPERIMENT

We have used the Random Forest implementation of the WEKA tool [7] to classify the data. WEKA is a Java package which contains machine learning algorithms for data mining tasks. We first converted the spambase data to .arff format that is supported by WEKA. In order to create different degrees of imbalance characteristics in the dataset, we first fixed the number of minority class instances and then varied the majority class instances in intervals of 100. All the instances are chosen randomly. As shown in Table 1 the spambase dataset has 1813 minority (spam) and 2788 majority (legitimate) class instances and we prepared different subsets of training data as follows: firstly _x the number of minority class instances to 100 and varied the number of majority class instances in intervals of 100 as (100, 100), (100, 200)...(100, 2400) so on. Then incremented the number of minority class instances to 200 repeating the first step as (200, 200)...(200, 2400) so on (1800, 1800), (1800, 1900)..(1800, 2400). This process is followed with minority class instances 300, 400, ... and so on. We generated these samples so that we could create 18 models. Some of these choices can bee seen in the graphs presented in Figure 2.

We also used a ten fold cross validation on training data for Random Forest. Using SMOTE we chose 100% increase of minority data and selecting the default neighbors as 5.

5. PERFORMANCE METRICS

In machine learning confusion matrix has been used significantly as a performance measure of classification algorithms. Confusion matrix shows the relationship between the actual class and the predicted class. It has four parameters true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Table 3 shows the confusion matrix defined for two classes in an imbalanced dataset. TP indicates the number of samples classified as true while they are true. True negative indicates the number of samples classified as false while they are false. False positive indicates the number of samples classified as true while they are false. False negative indicates the number of samples classified as false while they are true. Hence the measures FP and FN give the number of misclassified samples [12].

Table 3. Confusion matrix of imbalanced binary class

Class	Predicted (Minority)	Predicted (Majority)
Actual (Minority)	True Positive	False Negative
Actual (Majority)	False Positive	True Negative

From the confusion matrix four performance metrics can be derived: accuracy, sensitivity, specificity, and precision. Accuracy gives the percentage of correctly classified instances. For imbalanced data since the minority class is the class of interest it is represented as the positive class and the true positive rate is equal to the sensitivity. True negative rate or the accuracy of the majority class is equal to the specificity. For the experiment we plotted the graphs for the true positive rates computed using equation 1 [13]. For imbalanced data it is desirable to have a high true positive rate while maintaining reasonable true negative rates.

$$TP_{rate} = \frac{TP}{TP + FN} \tag{1}$$

The true positive rate effects the majority of these measures directly: accuracy and precision and hence the proposed model considers the effect of imbalance data on the true positive rate.

6. MODELING PARAMETERS

We modeled the relationship between the classification rate (true positive rate) using equation 1 and the ratio of minority class size (m_1 and majority class size m_2. Hence m_1 and m_2 are part of the set of modeling parameters. In this set up, we generated several linear models ($y = mx + c$) that are defined by the slope parameter m and the intercept parameter c.

Table 4. Breakpoints of True positive rates using Random forest with and without using SMOTE for spambase dataset

Minority class instances	Majority class instances - TP breakpoint using RF	Ration of minority to majority class	Majority class instances - TP breakpoint using RF and SMOTE	Ratio of minority to majority class
100	151	0.662251656	263	0.380228137
200	273	0.732600733	480	0.416666667
300	373	0.804289544	725	0.413793103
400	445	0.898876404	1006	0.397614314
500	635	0.787401575	1132	0.441696113
600	715	0.839160839	1276	0.470219436
700	913	0.766703176	1580	0.443037975
800	962	0.831600832	1830	0.43715847
900	1046	0.86042065	2100	0.428571429
1000	1195	0.836820084	2400	0.416666667
1100	1390	0.791366906	No breakpoint	
1200	1432	0.837988827	No breakpoint	
1300	1708	0.761124122	No breakpoint	
1400	1855	0.754716981	No breakpoint	
1500	1775	0.845070423	No breakpoint	
1600	1975	0.810126582	No breakpoint	
1700	2098	0.81029552	No breakpoint	
1800	2340	0.769230769	No breakpoint	

7. RESULTS AND FINDINGS

We plotted the true positive rates using both Random Forest with and without SMOTE by varying the degree of class imbalance in the dataset. These plots are presented in Figure 2 and in these plots the number of majority class instances are plotted in X axis and the true positive rate for the corresponding majority class instances is plotted in Y axis. Each figure belongs to a fixed minority class. For example, Figure 2(a) is for the minority class with 100 instances, Figure 2(b) is for the minority class with 300 instances and so on. In these plots, thick-lines represent the results of Random Forest and dashed-lines represent the results of Random Forest with SMOTE. In addition, the blue lines represent the results of majority class and red lines represent the results of minority class. From these plots, we noted the point of intersection or the breakpoint where both the minority and majority classes show the same true positive rates. The majority classes instances corresponding to these points of intersections are listed in second column (RF) and fourth column (RF with SMOTE) of Table 4 respectively. We also listed the ratio between the minority and majority classes at that instances in third column (RF) and fifth column (RF with SMOTE)) of this table. We can observe that the average ratio of minority to majority class instances is 0.8 for Random Forest and 0.42 for RF with SMOTE. This shows that SMOTE performs significantly better than Random Forest for imbalanced data as expected. It helps our modeling objectives.

Table 5. True positive line coefficient and intercepts for spambase dataset

Slope when RF is used	Intercept when RF is used	Slope when RF and SMOTE are used	Intercept when RF and SMOTE are used
-0.00013813	0.82057971	-7.08E-05	0.926758893
-0.000104447	0.912737154	-4.91E-05	0.963136646
-7.59E-05	0.935910032	-4.00E-05	0.976982684
-5.36E-05	0.931593074	-2.78E-05	0.975990602
-4.12E-05	0.938344361	-2.10E-05	0.977394737
-3.92E-05	0.951798246	-2.07E-05	0.983096549
-3.69E-05	0.961879208	-1.45E-05	0.98012605
-3.81E-05	0.971397059	-1.82E-05	0.989161765
-3.35E-05	0.971326797	-1.47E-05	0.98751918
-2.58E-05	0.971035714	-1.66E-05	0.993857143
-2.56E-05	0.973786214	-1.33E-05	0.992842158
-1.66E-05	0.96010989	-9.19E-06	0.986555458
-2.54E-05	0.982854581	-1.34E-05	0.996566434
-1.54E-05	0.969435065	-7.99E-06	0.988253247
-2.12E-05	0.981951515	-1.76E-05	1.008555556
-2.70E-05	0.997708333	-6.21E-06	0.988478423
-1.39E-05	0.973865546	-1.40E-05	1.003665966
-1.61E-05	0.983511905	-2.30E-06	0.982772487

We modeled the true positive curves for each minority class as a straight line that describes the relationship between the true positive rate and the majority class instances corresponding to a minority class. The slopes (*m*) and intercepts (*c*) are in the first and the second columns for RF and the third and the fourth columns for RF with SMOTE respectively. We found that the average slope of the line using RF is -0.0000415 and the average y intercept is 0.954 whereas for RF with SMOTE it is -0.000021 and 0.983 respectively. The slope of the line indicates that SMOTE is consistent with varying degrees of imbalances, the y intercept indicates that the true positive rate is higher when using SMOTE. It is also observed that the true positive rates of the majority class is slightly lesser when using SMOTE compared to random forest without SMOTE. This is natural as using SMOTE the number of instances of minority class is doubled so the classifier is more biased towards minority class if the majority class instances is fewer. Also as the number of minority class instances is increased the gap between the performance of RF with and without SMOTE is reduced. We plotted these average models for RF and RF with SMOTE in Figure 3 where the red line corresponds to the model associated with RF and blue line corresponds to the model associated with RF with SMOTE. The x-axis in this figure represents the ratio between majority and minority classes. This model is useful for predicting true positive rates when RF and RF with SMOTE are applied to big data classification where the imbalanced data is problematic.

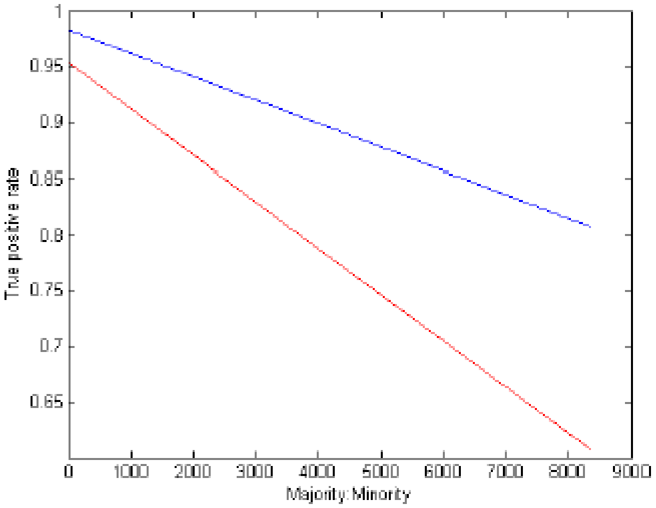


Figure 3. Imbalanced data with two classes

Figure 3, for example, provides the following information: when the imbalanced ratio between majority and minority classes in *u* : *u*, where *u* is large, then the true positive rate is about 0.95 for RF and 0.98 for RF with SMOTE; when this imbalanced ratio is 3000*u* : *u*, the true positive rate is about 0.84 for RF and 0.93 for RF with SMOTE; and when the ratio is 8000*u* : *u*, the true positive rate is 0.63 for RF and 0.82 for RF with SMOTE.

8. CONCLUSION

This research work shows that linear models that describe the relationship between true positive classification rate and the imbalanced ratio between the majority and minority classes can be generated using an empirical study with imbalanced datasets and classification techniques. In addition, an average linear model can be generated as a predictor to estimate the true positive classification rate for a particular imbalanced class ratio. The linear models that we fit can help in the development of new spam filter algorithms. Although the empirical study is conducted with spambase dataset, RF, and RF with SMOTE, it can be applied to other imbalanced datasets and classification techniques with SMOTE to develop linear models. The parameters of the model will further be studied in future to understand under which conditions the linear model works better. Finally, the proposed models can be used to determine the domain size when multi-domain classification techniques are explored for a big data classification.

9. REFERENCES

[1] R. Akbani, S. Kwek, and N. Japkowicz. "Applying support vector machines to imbalanced datasets." Machine Learning: ECML 2004. Springer Berlin Heidelberg, pp. 39-50, 2004.

[2] K. Bache, and M. Lichman. UCI machine learning repository. http://archive.ics.uci.edu/ml, 2013.

[3] D. Benyamin. "A gentle introduction to random forests, ensembles, and performance metrics in a commercial system." http://citizennet.com/blog/2012/11/10/random-forests-ensembles-and-performance-metrics.

[4] L. Breiman. "Random forests." Machine Learning, vol, 45, no. 1, pp. 5-32, 2001.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. "SMOTE: synthetic minority oversampling technique." arXiv preprint arXiv:1106.1813, 2011.

[6] C. Chen, A. Liaw, and L. Breiman. "Using random forest to learn imbalanced data," University of California, Berkeley, 2004.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. "The WEKA data mining software: an update." ACM SIGKDD explorations newsletter, vol. 11, no. 1, pp. 10-18, 2009.

[8] H. He, and E. A. Garcia. "Learning from imbalanced data." IEEE Transactions on Knowledge and Data Engineering, vol. 21, no. 9, pp.1263-1284, 2009.

[9] G. E. Hinton, S. Osindero, and Y. W. Teh. "A fast learning algorithm for deep belief nets." Neural computation, vol.18, no. 7, pp. 1527-1554, 2006.

[10] T. M. Khoshgoftaar, M. Golawala, and J. V. Hulse. "An empirical study of learning from imbalanced data using random forest." 19th IEEE International Conference on Tools with Artificial Intelligence, 2007, vol. 2, pp. 310-317, 2007.

[11] F. Provost. "Machine learning from imbalanced data sets 101." in Proceedings of the AAAI'2000 workshop on imbalanced data sets, pp. 1-3, 2000.

[12] D. Yao, Dengju, J. Yang, and X. Zhan. "An Improved Random Forest Algorithm for Class-Imbalanced Data Classification and its Application in PAD Risk Factors Analysis." Open Electrical and Electronic Engineering Journal, vol. 7, no. 1, pp. 62-70, 2013.

[13] Weiss, Gary M., and Foster Provost. "The effect of class distribution on classifier learning: an empirical study." Rutgers Univ (2001).