A Two-Way Sampling based on K-Means for Solving Class Imbalance Problem

Ms. Palle Sumana[1], Ms. M Sri Bala[2], Dr. K. S. M. V. Kumar[3]
[1]Mtech CSE student, [2]Sr. Asst Professor, [3]Professor,
Dept of CSE, LBRCE, LB Reddy Nagar, Mylavaram, Andhra Pradesh ,India

Abstract - The class imbalance problem has become more and more difficult to solve in the modern era due to the vast upsurge in the data which is produced every second. In the real world situations, the data-sets are generally composed in such a way that the classification categories are not equally distributed, which means there are relatively less irregular examples. And the expense involved in misclassifying the irregular data sample as regular is very high. This paper studies the imbalance classification problem and proposes a two way sampling approach based on K-means to obtain better classification results. It combines Synthetic Minority Oversampling Technique to avoid irregularities and shortage of samples.

Keywords - Imbalanced data, Smote, clustering, classification.

I. INTRODUCTION

Machine Learning is a field of computer science that uses statistical techniques to give the devices, the ability to learn in a progressive manner to perform better with a given data without being explicitly programmed. It grants us the ability to construct algorithms which can learn and predict data. It begins by building a model from the given inputs. It is closely related to computational statistics in which the target is to develop prediction by using computers. It can also be combined with data mining but it is more of analyzing data sets and comes under unsupervised learning.

Classification is one of the main problems in machine learning, where we need to identify which set of categories a new observation belongs based on the training set whose membership (association) is known. And a dataset is said to be imbalanced when the class distribution is varying. Since there has been a major evolution in the technology that we use today the amount of data has also increased exponentially. Therefore it has become difficult to address these issues with the present approaches of data mining. By combining with the machine learning approaches there is a scope of reducing the class imbalance problems.

II. THE PROBLEM

A data set is believed to be imbalanced when the classes are unequally represented. They are described as two classes, majority class and minority class. The majority class is class which consists of more number of instances (also known as negative class) and the other class which has relatively less number of instances is a minority class (also known as positive class). These incidents can be mainly seen in fraud detection, oil spills, medical diagnosis etc. It is a problem because in extremely rare cases, even if the classifier is trained such that it yields 99.9% accuracy on the test set but the output is completely different and it is not classified. This occurs because there are only a handful of instances which contain the same irregular categories compared to the thousands of other samples which have the similar characteristics. So it simply labels them as no fraud detected or no disease. Therefore the misclassification rate is high. The answer to this problem is to perform oversampling and undersampling techniques, one sided learning etc.

III. SOLUTION FOR CLASS IMBALANCE

Previous Work - In 1997 Kubat and Matwin considered the original population of the minority class and undersampled the majority class selectively. The performance measure they used for the classifier was the geometric mean. The minority samples divided into noise, border samples, redundant samples and safe samples. Another solution was to classify the superimposed region of positive and negative classes as positive. In 2000 Japkowicz considered two resampling methods. One method was random resampling which is to randomly resample the smaller class until there were samples almost equal to the majority class. Another method was focused resampling in which only the positive instances which were on the borderline of the positive and negative classes were resampled. Random undersampling was another method which considered the majority class instances at random and they were resampled until the numbers matched the minority class numbers.

Another approach is the metacost approach in which the probability of each class for each instance was predicted and they are labeled again concerning the misclassification rates. And also various resampling techniques like random undersampling, random oversampling (with replacement), directed undersampling, directed oversampling and the combinations of these techniques were used for decoding the imbalance class problem.

IV. PROPOSED WORK

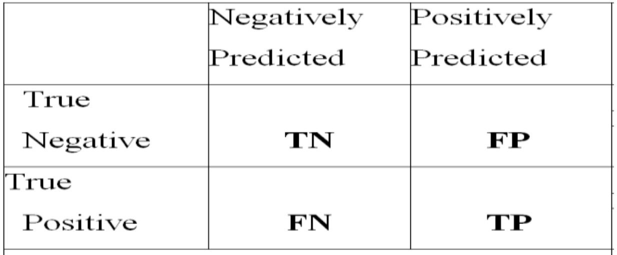
Depending on K-means, an under sampling method is prospected to remove the redundant samples by keeping the given sample as the center of the negative class samples. The difference between the sample under construction and its most adjacent neighbor is taken. The change is then multiplied by an arbitrary number between 1 and 0, and then aggregated to the sample under construction. This results in a random point on the segment of the line between two distinct instances, which changes the region of decision, of the positive class to become more generic. The costs of different classes are adjusted to avoid class imbalance.

V. SAMPLING BASED ON K MEANS IN TWO WAYS

Class imbalance involves between-the-class imbalance and within-the-class imbalance. The scenario in which the count of instances in the minority class differs from the count of instances in the majority class is the between-the-class imbalance. The scenario in which a class is made up of different sizes of clusters is called within-the-class imbalance problem. Both majority and minority classes are divided into various clusters using K-Means, and whichever instance is nearer to the center of the cluster is placed in the new dataset. The positive class is further divided into a pair of clusters using Synthetic Minority Oversampling Technique for oversampling which can ultimately obtain proper proportion between classes.

A. K-Means Algorithm: In data mining one of the most popular clustering techniques is the K-means clustering. The given n observations are broken into k clusters where every single observation is a part of the cluster having the nearest mean. These objects are assigned to the nearest collection based on the distance. At first the center of each cluster is chosen. Then based on the Euclidean distance the classification of the data is changed, according to this result the centers of the clusters are adjusted. This process is often mixed up with K-means because of its name and the entire result is actually based on the selection of the center of the initial cluster.

B. Performance Measures: This study mainly focuses on imbalance data classification. Compared to the majority-classes, the minority-classes have lower precision and recall values and the impact of accuracy is more on the major classes than the minority classes which makes it harder for the classifiers to classify the positive class data. To summarize the performance, the measures for this class imbalance problem are True-positive (TP), False-positive (FP), True-negative (TN), and False-negative (FN).Where True-positive means the various minority points correctly grouped, True Negative means various majority points correctly grouped, False Positive means the various minority data points wrongly classified and False Negative means the various majority points wrongly classified. This helps us to know the errors being made during computation.



	Negatively Predicted	Positively Predicted
True Negative	TN	FP
True Positive	FN	TP

Table 1: Error Matrix

The Precision of a class is its ability to return only relevant instances which is given by the number of true-positives divided by the entire elements which are addressed as the data points belonging to the positive class. And Recall can be understood as the ability of classification model to identify all relevant instances which is given by number of true-positives divided by the total number of elements that really belong to the true class.

$$Precision = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$Recall = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

The F_{measure} calculates the influence of recall and precision using Harmonic mean. It is defined as:

$$F_{measure} = \frac{(1+\beta^2) \times Recall \times Precision}{\beta^2 \times Recall + Precision}$$

The geometric mean is also used for calculating the result of positive class. It is given as:

$$G_{measure} = \sqrt{acc^{+} \times acc^{-}}$$

Where, accuracy+ and accuracy- are the accuracy values of positive class and negative class values respectively. They are given as,

$$accuracy^{+} = \frac{True\_Positive}{True\_Positive + False\_Negative}$$

$$accuracy^{-} = \frac{True\_Negative}{True\_Negative + False\_Positive}$$

The relation between Gmeasure and acc+ is nonlinear and they are directly proportional to each other, which means that most of the positive instances are wrongly classified.

Assume that the dataset is not balanced and consists of p classes and each class is of the size q1, q2, q3… The negative class is partitioned into k clusters using K-means where mean_value is the mean of class sizes. For minority class Smote oversampling is done. The pseudo code is framed as follows:

Input: Imbalanced dataset {C1, C2,.., Cp}, arithmetic mean_value of class size, number of minority class samples.

Output: Balanced dataset after resampling {C1’,C2’,…,Cp’}

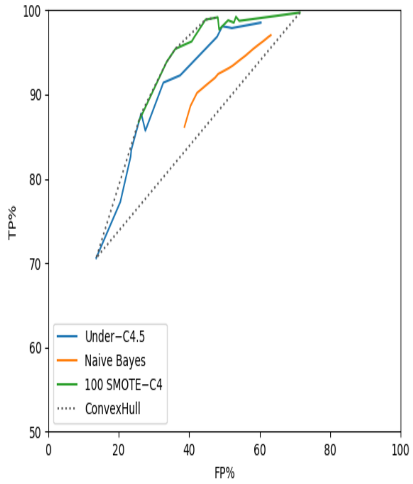
```
1. for i ← 1 to p
ni = class_size(Ci)
```

```
2. calculate mean of class_size
   $mean\_value = \sum_{i=1}^{p} \frac{q_1 + q_2 + q_3 + \cdots + q_i}{m}$

3. for i ← 0 to mean_value
If(class_size (Ci   >=   mean_value) )
It is a negative class perform K-means clustering and choose a cluster center for new dataset

Else

It is minority class

If smote_percent < 100

No_of_samples = $\frac{smote\_percent}{100} \times no\_of\_samples$

End if

4 Nearest_neighbour = k
   Initialize
   Sample_index[][],
   class_count[], synthetic_points[][]

For i ←1 to no_of_samples
Pick_neighbour at random = random(0,k)

While neighbourset != 0
Pick_neighbour == i

Pick_neighbour = random (0,k)

Difference = Ci[neighbourset[pick_neighbour]] – Ci[i]

Noise = random(1, no_of_features)

Synthetic_samples = Ci[i] + difference(1, No_of_features) x noise

End while

Samples += 1
Return synthetic_samples
End for
End
```

VI. EXPERIMENTAL RESULTS

To evaluate the result, the pima-indian diabetes data set has been chosen. The F1 value is high for random resampling and F1 value for minority class improves significantly as shown in the following. It is shown that the two-way sampling based on K means is better than the other methods. The other classification methods like Naive Bayes classifiers classify lesser number of minority samples and hence the imbalance problem does not reduce.

The C4.5 classifier shows better results than Naïve Bayes but as we see in the following graph K-means smote provide us more efficient balance in the data.



VII. CONCLUSION

The problem of Imbalance Data is present in the real world scenarios, especially for the minority classes. The two way sampling technique not only helps to classify the negative classes but also the minority classes and can improve accuracy in the classification of the positive class.

VIII. REFERENCES

[1]. H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering,2009, 21 (9):pp.1263–1284.

[2]. Y. Sun, A. K. C. Wong, M. S. Kamel, Classification of imbalanced data: a review, International Journal of Pattern Recognition and Artificial Intelligence, 2009, 23 (4):pp. 687-719.

[3]. J.V.Hulse, T. Khoshgoftaar. Knowledge Discovery from Imbalanced and Noisy Data [J].Knowledge and Data Engineering, 2009, 68(12): pp.1513-1542.

[4]. Z. M. Yang, L. Y. Qiao, X. Y. Peng. Research on datamining method for imbalanced dataset based on improved SMOTE [J].ACTA Electronica Sinica, 2007, 35(12): pp.22-26.

[5]. C. Drummond, R. Holte. Explicitly representing expected cost: An alternative to roc representation. In Proceedings of the sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2000, pp. 198–207.

[6]. G. E. A. P. A. Batista, R. C. Prati, M. C. Monard, A study of the behavior of several methods for balancing machine learning training data, SIGKDD Explorations,2004,6 (1): pp.20–29.

[7]. N. V. Chawla, K. W. Bowyer, L. O. Hall, W. P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, Journal of Artificial Intelligent Research, 2002, 16: pp. 321-357.

[8]. N. V. Chawla, L. O. Hall, K. W. Bowyer, W. P. Kegelmeyer. SMOTE: synthetic minority oversampling technique. Journal of Artificial Intelligence Research, 2002, 16: pp. 321-357.