Class Imbalanced classification using Genetic Approach

CH Mamatha
PG-Scholar, JBIET, India.

B Nageswara Rao
Associate Professor, JBIET, India.

Dr.P Srinivas Rao
Professor, HOD , JBIET, India.

Abstract – Class imbalanced problem is becoming crucial in the Artificial Intelligence (AI) and Machine Learning (ML) approaches, as increased usage of these techniques. The fuel for AI and ML techniques is data only, as the industry is moving towards the automation, the data is becoming crucial. Classification and prediction becoming crucial in these days, learning the model from these data will be biased. Accuracy of these models is important for predicting the new instances of the data. While doing this the false positives and false negatives will be biggest problems.

Index Terms – Class Imbalance, Machine Learning, Classification, Genetic Algorithm.

1. INTRODUCTION

The performance of machine learning algorithms is typically evaluated using predictive accuracy. Classification algorithms are biased to the majority class in the imbalanced dataset. Due to the down sampling will loses some critical information and it reduces the size of the dataset. The bad thing in the imbalanced dataset is that, the cost of missing a minority class is much higher than missing a majority class.

When the machine learning algorithms learns from imbalanced data, the learned model will be biased model towards the positive class. The existing techniques like oversampling and under sampling will balance the class. But these approaches suffers with the over fitting and under fitting. The smote algorithm is used to balance the data by using the oversampling and under sampling, by generating the synthetic data. These newly generated samples may not fit the model accurately, it will create the bias and variance tradeoffs. Because of these problems the learned model from the balanced data also may not predict the minority class accurately. While handling such class imbalances is crucial in learning the models. In the real time environment the availability of the data samples may not be balanced. This creates the lot of problems in predictive models of the imbalanced classes. To implement these problems need to be adopting new approaches for handling such kind of data for generating the new samples.

In the SMOTE algorithm, it generates the new synthetic data samples based on the existing data. It might be used for generating the new samples. Basically the smote algorithm uses the K-nearest neighbors algorithm for generating the new synthetic tuples. Majorly this algorithm is working on the principles of the nearest values. Classification algorithms tend to perform poorly when data is skewed towards one class, as is often the case when tackling real-world problems such as fraud detection or medical diagnosis. A range of methods exist for addressing this problem, including re-sampling, one-class learning and Farud Detections, Employee fraud, Payment fraud.

Providing an equal sample of positive and negative instances to the classification algorithm will result in an optimal result. Datasets that are highly skewed toward one or more classes have proven to be a challenge. Resampling is a common practice to address the imbalanced dataset issue. The following approaches are used to handle the class imbalanced data.

i. Random under-sampling – Reduce majority class to match minority class count.

ii. Random over-sampling – Increase minority class by randomly picking samples within minority class till counts of both class match.

iii. Synthetic Minority Over-Sampling Technique (SMOTE) [3] – Increase minority class by introducing synthetic examples through connecting all k (default = 5) minority class nearest neighbors using feature space similarity (Euclidean distance).

2. RELATED WORK

Big data has become an important topic worldwide over the past several years. Among many aspects of the big data research and development, imbalanced learning has become a critical component as many data sets in real-world applications are imbalanced, ranging from Internet, finance, social network,

to medical and health industry. In general, the imbalanced learning problem is concerned with the performance of machine learning algorithms in the presence of underrepresented data and severe class distribution skews. Due to the inherent complex characteristics of imbalanced data sets, learning from such data requires new understandings, principles, algorithms, and tools to transform vast amounts of raw data efficiently and effectively into information and knowledge representation.

A. Bias and Variance trade off

A fundamental problem with supervised learning is the bias variance trade-off. Ideally a model should have two key characteristics.

1. Sensitive enough to accurately capture the key patterns in the training dataset.

2. It should be generalized enough to work well on any unseen datasets.

Reddy et.al discussed about the handling of imbalanced data in anomaly detection using the robust data models [4], in this they combine the data approach and algorithm perspective and achieved good results. In this they used the class balancing approach using the smote approach. For classification ensemble methods are used. The combination of these two has enhanced the detection rate of the model.

Machine learning and data mining approaches are suffers with the class imbalanced problem. The skewed class is more biased than the other classes. In the real time environment the data available is highly imbalanced one. In the history of data mining many researchers contributed towards the class imbalanced problem [5].

When class contribution towards the decision making many of the methods deploy the statistical approaches. Learning from the imbalanced data is a big challenge in these days. There are many variants of solutions are existed for the imbalanced data, but they are not generating accurate solutions to the real world problems based on the existence of the data impurities. Imbalanced data is not the impurity of the data, we can overcome these problems by using the different approaches in the real time, these are tried to solve this problems.

i. Focus on the structure and nature of examples in minority classes in order to gain a better insight into the source of learning difficulties.

ii. Develop methods for multi-class imbalanced learning that will take into account varying relationships between classes.

iii. Propose new solutions for multi-instance and multi-label learning that are based on specific structured nature of these problems.

3. METHODOLOGY

The main aim this paper is to overcome the problems of class skewedness, machine learning approaches are not able bear these problems.

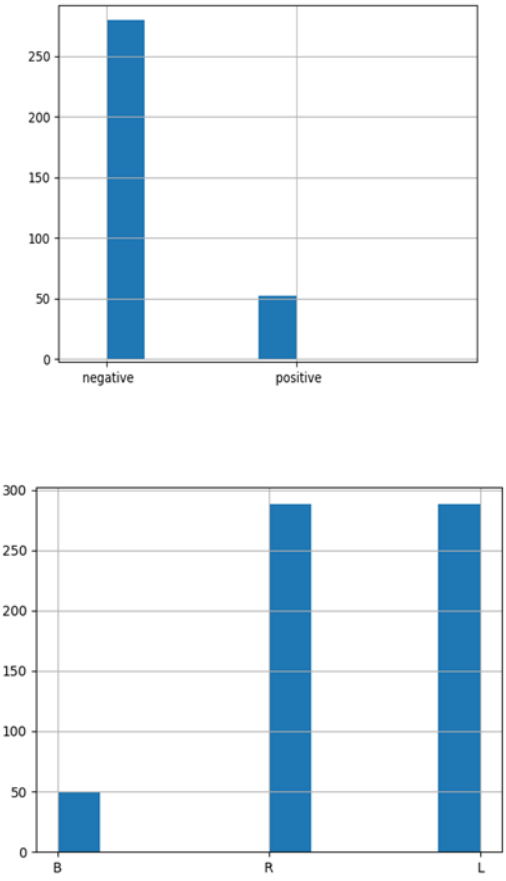

Figure 1: The class comparison for the ecoil and balance dataset

The class imbalance in the many datasets differs nearly more than 90% differs in between majority and minority class. some of the datasets may be binary classes and others may be multiclass datasets, which are imbalanced in multiclass are differs in distribution among the classes. As shown in figure 1, the class imbalances are shown. It will be differ for the different datasets.

Class imbalanced problem [2] can be handled in two ways. Many of the researchers assumed that class imbalanced is data problem, but few of the researchers solved these problems by changing the algorithm perspective and find the solutions for class imbalanced problems. To handle the imbalanced data the different approaches are used to resolve the issues. These can be categorized based on the how will handle the data or the

method used to handle the imbalanced data. Basically these are used to handle either the data or the approach. Many of the researchers contributed in these two directions to solve the class imbalanced problem.

1. Data-level Approach to class imbalance
    a. Under sampling
    b. Oversampling
2. Class_weight is your friend
    a. RF(Random Forest)
    b. SVM
    c. AdaBoost

Among these the data level approach itself is not suitable for handling the imbalanced problem. To overcome these problems we need to identify the appropriate solutions towards the impact of the problems and give the hybridized solutions.

Challenges

a. Another vital challenge is connected with the availability of class labels.
b. Most of existing works assume that we deal with a binary data stream for which the relationships between classes may change over time
c. In many real-life streaming applications (like computer vision or social networks) the imbalance may be caused by reappearing source.
d. Using characteristics and structure of minority class is a promising direction for static imbalanced learning

The goal is to minimize the cost of misclassification Synthetic samples of minority class can be generated as follows:

1. Consider the minority class feature vector and calculate the difference between with its nearest neighbors.

2. Multiply the difference by a random number between 0 and 1, and add it to the feature vector under consideration.

3. This causes the selection of a random point along the line segment between two specific features.

Based on the feature importances derived from this analysis, we have already switched off certain fraud alerts.What features are influencing the fraud score most; we provide a fraud score from 0 to 1.

4. RESULTS

The experiments are conducted for the different datasets collected from keel [12] which are imbalanced datasets

Table 1: Imbalanced datasets

S. No	Dataset Name	Features	Instances	% Positive instances (Minority Class)	% Negative instance (Majority class)	Class
1	Balance	4	625	14.53	85.47	Multi-class
2	Ecoli2	7	336	15.48	84.52	Binary
3	Wisconsin	9	683	34.97	65.03	Binary
4	Bank [13,15]	17	45211	11.52	88.47	Binary

The imbalanced data misses the minority class and the accuracy of the model will be crucial. In this article we compare the results of the different models based on the following approaches

1. Imbalanced data
2. Oversampled data
3. Under sampled data
4. Balanced data

The results of the different models are estimated on the different performance measures, which are listed below.

Table 2: Confusion Matrix for Performance metric

	Positive	Negative
Positive	Positive Identified as Positive (TP)	Negative Identified as Positive (FP)
Negative	Positive Identified as Negative (FN)	Negative Identified as Negative (TN)

1. $Accuracy = \frac{Tp+Tn}{Tp+Tn+Fp+Fn}$

Accuracy of the model will be used to predict the total percentage of correctly classified and identified objects from the total samples space.

2. $\text{Sensitivity(Recall)} = \frac{\text{Tp}}{\text{Tp+Fn}}$

It gives the positive results which are region of interest are identified correctly from the total positive samples.

3. $Precision = \frac{Tp}{Tp+Fp}$

The precision gives us the positive rate, the positive objects from the model classified positive objects.

4. $\text{F} - \text{measure} = \frac{2\times\text{precision}\times\text{recall}}{\text{precision+recall}}$

It is the mean of the precision and recall.

In this paper the different models are used to estimate the results for the different flavors of the imbalanced dataset with various performance metrics. The following approaches are used to estimate the results of the models for classification.

1. Support vector machines[1]
2. Random Forest.
3. K nearest neighbors

Table 3: The Results of the three different approaches for various techniques

S.No	Accuracy of the model			
	Imbalanced	Over sampling	Under sampling	Balanced
Support vector machines	81.44	72.13	67.51	83.62
Random Forest.	81.44	86.58	80.35	81.44
K nearest neighbors	84.8	77.75	75.32	79.61

The experiments are conducted for the three different techniques and for the various approaches of the data sampling techniques and accuracy is evaluated.

5. CONCLUSION

The imbalanced data is always skewed, which is from the different approaches and different models. These are verified with different approaches. The imbalanced data is suffers with bias and variance while fitting into the model. The accuracy of the model is varied for one technique to another and compared.

REFERENCES

[1] Variable importance in nonlinear kernels (VNK): Classification of digitized Histopathology. S. Ali, G Lee,
[2] Krawczyk, B. Prog Artif Intell (2016) 5: 221. https://doi.org/10.1007/s13748-016-0094-0
[3] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002).
[4] Ravinder Reddy R., Ramadevi Y., Sunitha K.V.N. (2016) Robust Data Model for Enhanced Anomaly Detection. In: Satapathy S., Bhatt Y., Joshi A., Mishra D. (eds) Proceedings of the International Congress on Information and Communication Technology. Advances in Intelligent Systems and Computing, vol 439. Springer, Singapore
[5] He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Trans. Knowl. Data Eng. 21(9), 1263–1284 (2009)
[6] He, H., Ma, Y.: Imbalanced Learning: Foundations, Algorithms, and Applications, 1st edn. Wiley-IEEE Press, New York (2013)
[7] Japkowicz, N., Stephen, S.: The class imbalance problem: a systematic study. Intell. Data Anal. **6**(5), 429–449 (2002)
[8] Wei, W., Li, J., Cao, L., Ou, Y., Chen, J.: Effective detection of sophisticated online banking fraud on extremely imbalanced data. World Wide Web **16**(4), 449–475 (2013)
[9] Woźniak, M.: Hybrid Classifiers—Methods of Data, Knowledge, and Classifier Combination. In: Studies in Computational Intelligence, vol. 519. Springer, Berlin (2014)
[10] Xu, R., Chen, T., Xia, Y., Lu, Q., Liu, B., Wang, X.: Word embedding composition for data imbalances in sentiment and emotion classification. Cogn. Comput. **7**(2), 226–240 (2015)
[11] Kamarulzalis A.H., Mohd Razali M.H., Moktar B. (2018) Data Pre-Processing Using SMOTE Technique for Gender Classification with Imbalance Hu's Moments Features. In: Saian R., Abbas M. (eds) Proceedings of the Second International Conference on the Future of ASEAN (ICoFA) 2017 – Volume 2. Springer, Singapore
[12] Fernández, A., García, S., del Jesus, M. J., and Herrera, F. 2008. A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. Fuzzy Sets and Systems 159, 18 (Sep. 2008), 2378-2398.
[13] https://archive.ics.uci.edu/ml/datasets/Bank+Marketing
[14] Lee, J. & Park, K. Pers Ubiquit Comput (2019). https://doi.org/10.1007/s00779-019-01332-y
[15] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

Author

CH Mamatha is a PG Scholar doing research in the machine learning for doing the imbalanced classification. Interested in anomaly detection and machine learning techniques. Completed her B.Tech in computer Science and Engineering.