Description of the Problem

A multi-class classification problem with class imbalanced data.

Literature Review

**Class Imbalance:**

* 1. Random Under sampling

Random under sampling is a technique that randomly eliminates observations from the majority class to match the number of observations in the minority class. Eliminating random observations may result in the loss of useful information. As mentioned by S. Wang [2], under sampling “suffers from performance loss of the majority class”, particularly when the discrepancy between samples is large. Given the size of the dataset and the small number of classes in the minority class, this method is unlikely to produce an accurate model.

* 1. Random Oversampling

Random oversampling is a technique that randomly replicates samples in the minority class to match the number of samples in the majority class. Unlike under sampling, there will be no information loss, however, the replication of samples in the minority class may result in an overfitted model. As mentioned by J.A. Saez [3], oversampling “may lead to increase in difficulty of the classification task”.

* 1. Synthetic Minority Oversampling Technique (SMOTE)

SMOTE is an oversampling algorithm that randomly picks a sample from a minority class and finds the k nearest neighbors for this sample. Synthetic samples are then added between the picked sample and its nearest neighbors and the process repeats for each class until the data is balanced. The synthetic samples will be different to the original samples and since they are close with the feature space on the minority class, new information is added to the data, resolving the issue of overfitting in random over sampling. However, as mentioned in T. Zhu [4], SMOTE is not very effective for high dimensional data as additional noise is introduced into the dataset which may affect the final accuracy of the model.

* 1. Cluster-based oversampling

Cluster-based oversampling is a technique that applies the K-means clustering algorithm independently to minority and majority instances to identify clusters in the dataset. Clusters are then oversampled until clusters of the same class have equal samples and class sizes are equivalent. The main drawback of this method is the potential for an overfitted model as there will be many repeated samples in the minority class [5].

* 1. SMOTE-Tomek

SMOTE-Tomek is a technique that combines the oversampling from SMOTE and under sampling from Tomek Links. The general process involves using SMOTE to create some synthetic data, and then choosing some random data from the majority class and removing them if the data are identified as Tomek links. This method has been used in practice with success due to its ability to handle class imbalance issues in data “by improving the predictive performance of classifiers” [6].

* 1. Ensemble techniques

An alternative approach to class imbalance is to consider ensemble techniques rather than data preprocessing. Ensemble models combine multiple models to improve and adapt models to imbalanced data. The diversification of models allows us to achieve a better overall fit. This technique has been used extensively for classifying imbalanced data in practice [7,8].

Challenges

One of the largest challenges we faced when building our algorithm was the class imbalance problem itself. The significant disparity between the highest and lowest number of samples between classes meant that we could not perform standard multi-class classification techniques on the given data. Two main approaches were explored to counter the class imbalance problem. The first approach involved tackling the problem on a data level. The main method used was resampling the training data, and techniques such as random under sampling, SMOTE and cluster based over sampling were employed. In this way, the training data would be resampled to eliminate the class imbalance problem, and then used to train standard machine learning models. The second approach involved modifying existing models to incorporate the class imbalance in the data. Adaptive boosting and gradient tree boosting techniques were explored in detail.

References

[1] R. Batuwita and V.Palade. “Efficient Resampling Methods for Training Support Vector Machines with Imbalanced Datasets”, The 2010 International Joint Conference on Neural Networks (IJCNN), 2010, oo. 1-8, doi: 10.1109/IJCNN.2010.5596787.

[2] S. Wang and X. Yao, "Multiclass Imbalance Problems: Analysis and Potential Solutions," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 42, no. 4, pp. 1119-1130, Aug. 2012, doi: 10.1109/TSMCB.2012.2187280.

[3] J.A. Saez, B. Krawczyk, M. Wozniak. “Analyzing the oversampling of different classes and types of examples in multi-class imbalanced datasets” in Pattern Recognition, vol. 57, pp. 164-178, September. 2016.

[4] T. Zhu, Y. Liu, Y. Liu. “Synthetic minority oversampling technique for multiclass imbalance problems” in Pattern Recognition, vol. 72, pp. 327-340, December. 2017.

[5] T. Zhu, Y. Liu, Y. Liu. “Improving interpolation-based oversampling for imbalanced data learning” in Knowledge-Based Systems, vol. 187, 104826, January 2020.

[6] Y. Sanguanmak and A. Hanskunatai, "DBSM: The combination of DBSCAN and SMOTE for imbalanced data classification," 2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2016, pp. 1-5, doi: 10.1109/JCSSE.2016.7748928.

[7] S. Oh, M. S. Lee and B. Zhang, "Ensemble Learning with Active Example Selection for Imbalanced Biomedical Data Classification," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 8, no. 2, pp. 316-325, March-April 2011, doi: 10.1109/TCBB.2010.96.

[8] G. Haixiang, L. Yijang, L. Yanan, L. Xiao, L. Jinling. “BPSO-Adaboost-KNN ensemble learning algorithm for multi-class imbalanced data classification” in Engineering Application of Artificial Intelligence, vol. 49, pp. 176-193, March 2016.