A Cost-Sensitive Learning Approach for Handling Imbalanced Data in Decision Trees

Aidan Lee and Sami Saleh

1/16/24

Yilmaz ML Pd.4

Quarter 2 Project

Abstract

Class imbalance is a largely prevalent issue when it comes to classification. There are different ways to combat this problem. One way is through cost-sensitive learning, which is a method that assigns different "costs" to misclassification errors. In this paper, we utilize a cost-matrix approach alongside a cost-sensitive Gini Index algorithm for Decision Trees in order to address this model's issues with imbalanced data. We found that our algorithm does little to improve Decision Trees' handling of imbalance data, but there is potential for improvement in testing different values for the cost matrix.

Introduction

Imbalanced datasets can create many issues in machine learning classification algorithms, particularly when the misclassification varies largely across classes. For example, when diagnosing medical cases, false negatives can lead to much more severe consequences when compared to false positives. This would lead to a patient thinking that they don't have cancer, when they in fact do have cancer. Normally, decision trees are often biased towards the majority class since it splits nodes in a way that reduces overall errors. Since the majority class is more dominant, the algorithm may lean more towards it.

In our paper, we will explore how we can use cost-sensitive learning in decision trees to improve the classification accuracy on imbalanced datasets. We will be using a cost matrix to minimize the misclassification costs. Our primary dataset is the Credit Card Fraud Detection dataset. It contains transactions made by credit cards in September 2013 by European cardholders. The input to our algorithm are the different instances with the transaction features (anonymous to protect privacy) and we used decision trees to output a binary classification of 0 (non-fraudulent) and 1 (fraudulent).

Related Work

There has been significant research done on how we can address imbalanced datasets in machine learning. However, cost-sensitive learning is a relatively new field with only small

amounts of research into its applications on decision trees and other machine learning models. Nguyen et. al. found that cost-sensitive learning performs just as well and in some cases better than other standard models such as SVMs. In addition, Mienye and Sun tested cost-sensitive learning on different models, including decision trees, and found overall precision rates and recall rates in the 70s. However, they applied a different form of a cost-sensitive learning algorithm than us and on different datasets as well, so we expect different results than these two researchers.

Dataset/Features

Our dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, with the positive class (frauds) accounting for only 0.172% of all transactions. There are 30 features, 28 of which are represented by "V1, V2, …V28", and these features are normalized already and their information is unknown to preserve the privacy of the credit card holders. There is also a "Time" feature which states how long after the first instance the transaction occurred. Finally, there is an "Amount" feature containing how much the transaction was. The class attribute was "Class" with a 0 representing a non-fraudulent transaction and a 1 representing a fraudulent transaction. The features are not that important to us, as it is the class imbalance that we want to focus on and how to improve the decision trees algorithm for imbalance data. For preprocessing, we did the following:

1. Removed the "Time" attribute from the dataset as it doesn't provide any valuable information
2. Discretized each of the V1-V28 features and the "Amount" feature as well into three bins each.
3. Converted all data types from Numerical to Nominal using WEKA NumericToNominal function.
4. Conducted a 70-30 train-test split leaving

Methods

For our research project, we opted for using a Decision Tree approach, as this model often struggles with imbalanced data. A Decision Tree is a supervised learning algorithm used for both classification and regression tasks (in our case classification). It works by iteratively splitting the dataset into subsets based on feature values, creating a tree-like structure where each node represents a feature, each branch represents a decision, and each leaf node represents an output class. The goal of the algorithm is to find the splits that best separate the data according to a chosen heuristic, such as Gini impurity or Entropy. This process continues until a stopping condition is met, such as reaching a maximum depth or achieving pure leaf nodes.

Within the Decision Trees algorithm, we opted to use the gini impurity, given by the following equation:

$$\text{Gini} = 1 - \sum_{j} (p_j^2)$$ where p_j is the probability of class j.

The gini impurity measures the frequency at which any element of the dataset will be mislabelled when it is randomly labeled. Zero, the minimum value of the Gini index, indicates that the node is pure, so an optimal split is one with the smallest Gini Index. Our reason for choosing the Gini Index as opposed to Information Gain (Entropy), is that the logarithm computations in Information Gain are expensive especially when dealing with a large dataset as we are.

For our research project, we looked to test a unique form of the Gini impurity heuristic that incorporates cost-sensitive learning. We first initialized a cost-matrix that contains weights based on the misclassifications. A general method for class weighting is to utilize the inverse of the class distribution of the dataset. In our case, since there were 284,315 instances with the 0 class and 492 instances with the 1 class (a ratio of approximately 577-1), we created our cost-matrix as the following:

0	**1**
500	**0**

Here 500 corresponds to the cost of classifying an instance with class 1 as class 0, and 1 corresponds to the cost of classifying an instance with class 0 as class 1. Note the main diagonal is zeros because there should be no cost for correctly classifying instances.

Our main algorithm that we are testing is a variant of the Gini Index that incorporates cost-sensitive learning as follows:

$$\square\square\square\square\ \square\square\square\square\square\square\square\square\square\ \square\square\square\square = \sum_{\square=1}^{\square} \square_{\square} * (\sum_{\square \neq \square} \square\square\square\square(\square, \square) * (1 - \square_{\square}))$$

Where $\square\square\square\square(\square, \square)$ is the cost of misclassifying an instance of class i as class j

This formula ensures that the decision tree penalizes misclassifications of the minority class more heavily by incorporating the cost matrix, so if the cost of misclassifying the minority class is high, the Gini index will increase, encouraging the decision tree to prioritize correct classification of the minority class.

We will also test this algorithm on the same dataset after applying SMOTE (Strategic Minority Oversampling Technique) to achieve a better majority-minority class ratio (20:1). This is more realistic than the original dataset ratio (577:1), and we think this will lead to more appropriate results that could possibly be applied to real imbalanced datasets.

Results

Since we are dealing with imbalanced data classification, our main metrics are precision, recall, and f1 score. We are working with a very large and imbalanced dataset, so accuracy will not be sufficient for understanding the effectiveness of our model. We opted to go with Precision, Recall, and F1-Score as our metrics for evaluating the models.

$$\square\square\square\square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square} \qquad \square\square\square\square\square\square = \frac{\square\square}{\square\square + \square\square} \qquad \square 1 = \frac{2 * \square\square\square\square\square\square\square\square\square * \square\square\square\square\square\square}{\square\square\square\square\square\square\square\square\square + \square\square\square\square\square\square}$$

Here, TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives

Table 1: Metrics Results

Model Type	Precision	Recall	F1-Score
DT	0.713	0.755	0.733
CSL-DT	0.701	0.701	0.701
DT w/ SMOTE	0.974	0.862	0.915
CSL-DT w/ SMOTE	0.974	0.862	0.915

Overall, Decision Trees without our cost-sensitive learning algorithm seemed to perform just as well and even better in some cases. However, when incorporating SMOTE beforehand, both models performed equally well, suggesting that our algorithm does little to improve Decision Tree's ability to handle unbalanced data. However, since the F1-score for our cost-sensitive decision trees model is 0.915, this shows that there is potential for cost-sensitive learning to aid in classification of imbalance datasets. We do believe though that more research is needed into how changing the value of the cost can affect the results, and this needs to be tested on multiple imbalance datasets in the future.

It does seem that our model is overfitting the training data, and this is because the precision, recall, and f1 score are all significantly higher than in the testing data. This overfitting is likely a result of us setting a max depth in the decision tree model, cutting off more possible branches and thus having a tendency to overfit the training data and leaving out crucial information. There was not much we could do to mitigate this as the process time with this depth already took a substantial amount of time per run, but with a more efficient model we believe it could be more robust to this overfitting. With more time, we would like to test our algorithm on more imbalanced datasets and further research its potential.

Contributions

Abstract: Aidan

Introduction: Aidan

Related Work: Aidan

Dataset/Features: Aidan

Methods: Sami

Results: Sami

Conclusion: Sami

Slides: Aidan & Sami

Works Cited

Aznar, Pablo. "Decision Trees: Gini vs Entropy ⋆ Quantdare." *Quantdare*, 13 Dec. 2020, quantdare.com/decision-trees-gini-vs-entropy/.

Brownlee, Jason. "Cost-Sensitive Decision Trees for Imbalanced Classification." *MachineLearningMastery.Com*, 20 Aug. 2020, machinelearningmastery.com/cost-sensitive-decision-trees-for-imbalanced-classification/.

Mienye, Ibomoiye, and Yanxia Sun. "Performance Analysis of Cost-Sensitive Learning Methods with Application to Imbalanced Medical Data." *Informatics in Medicine Unlocked*, Elsevier, 3 Aug. 2021, www.sciencedirect.com/science/article/pii/S235291482100174X#bib28.

Thai-Nghe, Nguyen. *Cost-Sensitive Learning Methods for Imbalanced Data*, www.ismll.uni-hildesheim.de/pub/pdfs/Nguyen_et_al_IJCNN2010_CSL.pdf. Accessed 28 Jan. 2025.