ISSN: 2769-545X

Review Article

Advances in Machine Learning & Artificial Intelligence

Applying Minority Range to Gini Index to Handle Imbalanced Dataset in Decision Tree Classifiers

Ben Mathew* and Marius Silaghi

Ph.D. Student at FloridaTech, Melbourne, Florida, United States.

***Corresponding Author**
Ben Mathew, Ph.D. Student at FloridaTech, Melbourne, Florida, United States.

Submitted: 2023, Oct 06; **Accepted**: 2023, Dec 13; **Published**: 2024, Jan 08

Citation: Mathew, B., Silaghi, M. (2024). Applying Minority Range to Gini Index to Handle Imbalanced Dataset in Decision Tree Classifiers. *Adv Mach Lear Art Inte, 5*(1), 01-03.

Abstract

The Class imbalance Problem is a common problem in Machine Learning where the number of instances in one class is significantly lower than the other, this can lead to biased classification models where the majority class dominates and the minority class is mis- misclassified. Decision Tree Classifiers are commonly used for classification tasks due to their simplicity and in interpretability. However, the class imbalance Problem can negatively impact the performance of decision tree Classifiers [1]. In this paper, we discuss a new approach to training a decision tree classifier that is an improvement over a pre-existing approach. We also provide experiments to prove that the proposed method is an improvement over the preexisting metrics.

Keywords: Class Imbalance problem, Machine Learning, Classification

1. Introduction

Decision Trees are the most popular model of computation for the classification of data. The decision tree model can be built using various models like ID3 [2] and C4.5 algorithms. This algorithm applies a recursive partition method to construct the tree. However, these algorithms have been shown to have certain limitations towards imbalanced datasets. The types of datasets have the characteristic feature of having a different number of instances among the classes. In an imbalance, a class with a smaller number of instances is known as the minority class and a class with a larger number of instances is known as the majority class. The accuracy of the model to predict the minority class is an area of interest. This is especially true in situations where failing to predict a minority class correctly can have significant real-world implications such as in the medical field where one can fail to diagnose cancerous tissue. One solution presented by [3] is to consider entropy only in the vicinity of the minority class range when deciding on which attribute to split on during the induction step.

However, this paper uses entropy as a measure of impurity which is known to increase training time. In this paper, we propose combining the Gini index with the minority entropy approach presented in [2]. Which we hope will decrease training time without any impact on accuracy.

1.1 Background

We plan to read the following papers listed in the reference section namely [4-6] a popular technique discussed which is called the Synthetic minority sampling technique. The above method is an oversampling technique where the minority class is over-sampled by taking each of the minority class samples and introducing synthetic data into the training examples. Although the above techniques alleviate overfitting by introducing random sampling it does not take into consideration neighboring examples from other classes and thus can introduce additional noise. Additionally, the approach discussed in [4] which is Adaptive Synesthetic Sampling generates synthetic samples inversely to the density of examples in the minority class which suffers from the sample drawbacks as [5]. In Contrast, the method discussed in [paper] tries to modify the induction process for the decision tree classifier. During the induction step, the author uses entropy as a measure of information gain to evaluate all instances that lie in the vicinity of the minority class. One potential drawback of using such a method is that the training time might be greater than when we use other metrics like

Gini- index to train the model.

1.2 Problem Formulation

We are going to modify the standard decision tree algorithm described in [7]. In each iteration of the algorithm when choosing the attribute that best classifies the instances, we will use the Gini index combined with the approach described in [4]. To Elaborate further, the measure for best classification taken by [2] is to evaluate entropy over the instances that lie in the vicinity of the minority class. In our work, we will instead use the Gini index over the instances that lie in the vicinity of the minority class. Specifically, let spra (D) represent the set of all instances (from both minority and majority classes) for which the value of attribute "a" lies within the range of possible values of "a" in the minority class. Where D is the set of all training Examples. The definition of spra(D) can be given as follows: $spr_a(D) = \{i \in D | minproj_a(K) \leq proj_a(i) \leq maxproj_a(\text{l})\}$

1.3 Evaluation Criteria

We plan to the following dataset from the UCI repository. Wine dataset:

The wine dataset is a dataset that consists of continuous attributes. The attributes in the dataset are namely:

- Alcohol
- Malic acid
- Ash
- Alkalinity of ash
- Magnesium
- Total phenols
- Flavanoids
- Nonflavanoid phenols
- Proanthocyanins
- Color intensity
- Hue
- OD280/OD315 of diluted wines
- Proline

1.3.1 Breast Cancer Dataset: This dataset consists of real-valued attributes and these attributes include:

- radius (mean of distances from the center to points on the perimeter)
- texture (standard deviation of gray-scale values)
- perimeter area
- Smoothness (local variation in radius lengths)
- compactness (perimeter2 / area -1.0)
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry

1.3.2 Iris Dataset: This consists of real-valued attributes namely:

- sepal length in cm
- sepal width in cm
- petal length in cm
- petal width in cm
- class

1.4 Experimental Results

In our experiment we apply the Minority Gini Heuristic on the wine (?) and Ecoli Dataset (?). We then use the Accuracy as a measure to compare the accuracy of the hypothesis that is derived from using the Minority Gini to the Minority Entropy as well as Entropy (Which is the traditional heuristic used in the ID3 model). We compute the accuracy of all three decision trees based on the training set, test set, and validation set. The data from the above experiment is tabularized in the following tables

Training Time

Data-Set	Entropy	Minority- Entropy	Minority- Gini-Index
Iris	0.0816	0.0961	0.093
Wine	0.243	0.2638	0.2578
Breast	2.4092	3.2109	2.5092

Accuracy

The accuracy of the decision tree on the training dataset can be summarized in the following tables.

Training Data-Set

Data-Set	Entropy	Minority- Entropy	Minority- Gini-Index
Iris	1.0	1.0	1.0
Wine	1.0	1.0	1.0
Breast	1.0	1.0	1.0

Test Data-Set

Data-Set	Entropy	Minority- Entropy	Minority- Gini-Index
Iris	0.96	0.94	0.94
Wine	0.93	0.87	0.87
Breast	0.94	0.94	0.93

2. Results and analysis

From the above data represented in the tables, we can infer the following regarding the heuristic used to split the data at a given node

• Accuracy: From the above data we can make the following inference on each of the datasets.

• Training Dataset: n terms of training accuracy there is no significant inference that can be observed as we see a 100

• Test Dataset: We see that MinorityGini is comparable to minority Entropy in terms of accuracy on the test data.

• Training Time: In terms of training time, we see that Minority Gini is significantly faster than Minority Entropy on all datasets used in the experiment. It can also be observed that the Minority Gini is comparable to that of entropy in terms of training time.

3. Conclusion

Thus, from the above paper we introduce Minority Gini-Index as a measure to split a node in a decision tree induction process. Additionally, from our experiments on the various datasets used in our experiments, we have shown that the new heuristic is comparable to the original induction process which uses Entropy as a measure in terms of training time and accuracy. Additionally, it has shown to be an improvement over minority Entropy.

References

1. Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. Smote: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16:321–357.
2. Hssina, B.; Merbouha, A.; Ezzikouri, H.; and Erritali, M. 2014. A comparative study of decision tree id3 and c4. 5. International Journal of Advanced Computer Science and Applications 4(2):13–19.
3. Shen, Y.; Liu, H.; Wang, Y.; Chen, Z.; and Sun, G. 2016. A novel isolation-based outlier detection method. In PRICAI 2016: Trends in Artificial Intelligence: 14th Pacific Rim International Conference on Artificial Intelligence, Phuket, Thailand, August 22-26, 2016, Proceedings 14, 446–456. Springer.
4. He, H.; Bai, Y.; Garcia, E. A.; and Li, S. 2008. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In 2008 IEEE international joint conference on Neural Networks (IEEE World Congress on computational intelligence), 1322–1328. Ieee.
5. Dua, D., and Graff, C. 2019. Uci machine learning repository [http://archive. ics. uci. edu/ml]. irvine, ca: University of California, school of Information and computer science. IEEE Transactions on pattern analysis and machine intelligence 1(1):1–29.
6. Subhani, S.; Gibescu, M.; and Kling, W. 2015. Autonomous control of distributed energy resources via wireless machine-to-machine communication; a survey of big data challenges. In 2015 IEEE 15th International Conference on Environment and Electrical Engineering (EEEIC), 1437–1442. IEEE.
7. Michalski, R. S.; Carbonell, J. G.; and Mitchell, T. M. 2013. Machine learning: An artificial intelligence approach. Springer Science & Business Media.