Thomas Jefferson High School

Q2 Project Proposal

Student :
Ishan Ajwani
Logan Bradley

Teacher :
Dr. Yilmaz

December 12, 2024

Contents

1 Motivation

The K-Nearest Neighbors algorithm relies on majority voting amongst the nearest neighbors of an instance to determine its classification. However, this algorithm is inherently flawed in that it favors classes with higher representation in the dataset, leading to frequent misclassification of infrequent labels. This problem is particularly prevalent in domains such as medical diagnosis, where a large majority of patients do not test positive, yet false negatives can be extremely detrimental. Our aim is to address this bias by introducing new pre-processing and weighting techniques into the model training process, improving the accuracy of KNN classifiers with respect to false negatives.

2 Method

Our plan to improve upon the K-Nearest Neighbors algorithm is based on a modified pre-processing and weighting approach. There are four changes we aim to test. Firstly, we will utilize the Synthetic Minority Over-sampling Technique (SMOTE) to reduce the skew in non-uniform datasets and increase the frequency of minority classifications, which should strengthen the KNN model. Chawla et. al (2011) found significant improvements by implementing SMOTE, outperforming a method of simply under-sampling the majority class. The traditional KNN algorithm simply does a majority vote of the K nearest neighbors. However, the algorithm can be enriched by considering three additional factors. Firstly, local density weighting can be used: neighbors that are in a denser region are generally stronger indicators of a classification than those in sparsely populated regions. This notion can be utilized by increasing the weightage of neighbors in dense regions to have a larger impact on the final classification. Secondly, weighted averaging can be used to account for large skews. An arbitrary inversely proportional relationship can be constructed between a neighbor's classification's frequency in the dataset and its weightage. For example, a neighbor that is classified as a label which is extremely rare in a dataset will have a higher weightage in the final outcome than one which is classified as a common label. Lastly, weighted distancing can be utilized to further sophisticate the KNN algorithm. By factoring the distance of each neighbor from the instance to be classified, more accurate predictions can be made.

3 Intended Experiments

To evaluate what properties make the best KNN classifier, we will use a variety of datasets and differing models. To ensure that the modifications we implement are not decreasing the potency of the classifier on regular datasets, we will test the classifier on both uniform datasets and highly skewed datasets. In addition to varying datasets, we will evaluate the performance of separate improvements by comparing a baseline KNN, a KNN using weighted averaging, a KNN using weighted distancing, a KNN using local density weighting, and a KNN using all the methods combined. Each model will also be trained using standard pre-processing techniques and SMOTE. Evaluation metrics will extend beyond accuracy, as the main focus of this project is ensuring a low false negative rate. Instead, we will utilize confusion matrices and recall scores.

4 Sources

https://www.researchgate.net/profile/Bayan-Shawar/publication/262165219_Bayesian-Base
Instance_Weighting_Techniques_for_Instance-Based_Learners/links/56b88a2b08ae5ad3605f3
Bayesian-Based-Instance-Weighting-Techniques-for-Instance-Based-Learners.pdf
https://www.jair.org/index.php/jair/article/view/10302/24590