CPS844: Data Mining

Assignment 1

March 4, 2023

Professor Elodie Lugez

Mark_Belleza	501026638
Sujeeta Warraich	500918432

Introduction

This report presents a comprehensive analysis of the Adult dataset from the UCI Machine Learning Repository[1]. The choice of this dataset is motivated by its substantial size and the complexity inherent in its attributes, making it an exemplary candidate for evaluating a range of classification algorithms. The dataset is composed of 48,842 instances, each with 14 attributes, reflecting census data aimed at predicting whether an individual's income exceeds $50,000 annually.

The classification methods used in this study are Decision Tree, K-Nearest Neighbors (KNN), Naive Bayes, Support Vector Machines (SVM), and Artificial Neural Networks. These methods were selected due to their wide application and varying approaches to the learning process, which offers a comprehensive examination of the dataset's predictive capabilities.

In the subsequent sections, we will delve into each classification method and the outcomes obtained from their application to the Adult dataset. Special attention will be given to evaluating performance metrics such as accuracy and computational efficiency. Finally, the concluding segment will synthesize the results, comparing the methodologies to determine the most effective classifier in terms of prediction accuracy.

Problem

Calculate and analyze the accuracy of different classifiers using the provided dataset.

Classifiers Used and Results

Decision Tree Classifier

Accuracy using Decision Tree (‘gini’, max_depth = 10): **0.857976354982343 (86%)**
Accuracy using Decision Tree (‘gini’, max_depth = 5): **0.8493781667434362 (85%)**
Accuracy using Decision Tree (‘entropy’, max_depth = 10): **0.8527560264087211 (85%)**
Accuracy using Decision Tree (‘entropy’, max_depth = 5): **0.8464609243052357 (85%)**

The performance of the Decision Tree classifier on the Adult dataset is assessed using two different criteria for splitting nodes: the 'gini' index and 'entropy', also known as information gain. Additionally, two levels of tree depths are evaluated: a depth of 10 and a depth of 5. The highest accuracy is observed using the 'gini' index with a maximum depth of 10, achieving approximately 86%. This suggests that at this depth, the Decision Tree is able to capture the complexity of the data well without significant overfitting.

When employing 'entropy' as the criterion, the accuracies for both depths (10 and 5) are roughly 85%. This is slightly lower than the 'gini' index at a depth of 10, suggesting that for this particular dataset, the 'gini' index may be a more effective measure for quality of splits in the context of predicting income levels.

K Nearest Neighbor Classifier

Accuracy using K-Nearest Neighbors (n_neighbors = 100): **0.7922616305849839 (79%)**
The K-Nearest Neighbors (KNN) classifier's performance with the Adult dataset is consistent across a wide range of 'n_neighbors' values, from 10 to 10,000, maintaining an accuracy of approximately 79%. This indicates that the choice of 'n_neighbors' does not

significantly impact the model's ability to predict income levels above or below $50,000 within this dataset.

Naive Bayes Classifier

Accuracy using Naive Bayes: **0.7996315062183326 (80%)**

The Naive Bayes classifier's accuracy of approximately 80% on the Adult dataset is a notable result given the simplicity and assumptions of the algorithm. Naive Bayes operates under the assumption that the presence or the absence of a particular feature of a class is unrelated to the presence or the absence of any other feature, given the class variable. In the case of the Adult dataset, which includes attributes such as age, education, occupation, and race, among others, the Naive Bayes algorithm has managed to provide a relatively high accuracy. This could be due to the presence of a few highly predictive features that dominate the others, or it may suggest that the interdependencies between features are not strong enough to significantly diminish the classifier's predictive power.

Support Vector Machines (SVM)

Accuracy using SVM: **0.7950253339474896 (80%)**

An accuracy of 80% indicates that SVM was able to find a hyperplane that adequately separates the data into two classes: individuals earning above or below $50,000. This is achieved by maximizing the margin between the data points of the two classes, which is a fundamental principle of SVM known as the maximization of the decision margin.

Artificial Neural Network

Accuracy using Artificial Neural Network: **0.7956395149230957 (80%)**

The Artificial Neural Network's performance, with an accuracy of approximately 80% on the Adult dataset is on par with the other sophisticated classifiers like SVM and Naive Bayes. This performance indicates that the neural network has successfully captured the underlying patterns and relationships in the data to predict whether an individual earns more than $50,000 per year.

An accuracy of 80% suggests that the ANN was able to model the non-linearity in the data without overfitting. The success of the ANN on the Adult dataset also implies that the network architecture, including the number of layers and the number of neurons in each layer, was sufficiently well-specified to capture the essence of the data without being too simplistic or excessively complex.

Analysis

Out of all the 5 classifiers used, the Decision tree classifier seems to give the most accurate predictions, both in gini and entropy criterion. While the decision tree classifier reached an average of 85% accuracy, the 4 remaining classifiers could only get to the high 70s in accuracy.

In terms of the importance of each attributes, it seems that the attribute that hold the most importance to the classifiers is '**capital-gain**'. There is a noticeable decrease in accuracy by a few percentage across all classifiers when this attribute is removed. Furthermore

when three attributes along with the attribute ‘**education-num**’ are dropped, an observable change can also be seen in the accuracy result only in the decision tree classifier.

For example. when dropping the attribute: ['**capital-gain**']

Gives the following results:

Accuracy using Decision Tree (‘entropy’, max_depth = 10): **0.8271150007676954 (83%)**

Accuracy using K-Nearest Neighbors (n_neighbors = 100): **0.7663135267925687 (77%)**

Accuracy using Naive Bayes: **0.7747581759557808 (77%)**

Accuracy using SVM: **0.7686166129279902 (77%)**

Accuracy using Artificial Neural Network: **0.7686166167259216 (77%)**

As seen, there is an observable difference in accuracy across all the classifiers used. There is a 3% decrease in the decision tree, similar to the previous example, 2% decrease in K-Nearest Neihbors and 3% decrease in Naive Bayes, SVM and Artificial Neural Network. This displays the significance the attribute ‘**capital-gain**’ plays in this dataset and to the classifiers. This suggests that **'capital-gain**' is likely a strong indicator of an individual's economic status and, by extension, their overall income category.

Furthermore, when the following attributes are removed:

['workclass', 'occupation', 'education', '**education-num**']

Gives the following result:

Accuracy using Decision Tree (‘entropy’, max_depth = 10): **0.8301857822815907 (83%)**

Interestingly, only the decision tree classifieris affected in the accuracy result. Increasing or decreasing the max_depth does not seem to have an effect this result either showing its importance as an attribute in this dataset and to the decision tree classifier.

Conclusion

In conclusion, the analysis of the Adult dataset employing various classifiers reveals valuable insights into the predictive capabilities of each model. Among the classifiers examined, the Decision Tree classifier emerges as the most accurate, consistently achieving an average accuracy of 85% across different criteria. This highlights the strength of decision trees in capturing complex relationships within the dataset.

Furthermore, attribute importance analysis underscores the significance of 'capital-gain' in predicting income levels, with its removal leading to noticeable decreases in accuracy across all classifiers. This underscores its crucial role as a predictor of economic status and income categories within the dataset. Overall, this analysis provides valuable insights into the performance and behavior of different classifiers on the Adult dataset, offering guidance for further refinement and optimization of predictive models in similar socio-economic prediction tasks.

References

[1] Becker,Barry and Kohavi,Ronny. (1996). Adult. UCI Machine Learning Repository.

https://doi.org/10.24432/C5XW20