COMP551 MiniProject 1

Group 58: Saad Benslimane Dun Yuan Yuhongze Zhou

Abstract

In this project, we explored two datasets – `Adult` dataset and `Avila` dataset from UCI Machine Learning Repository. We firstly applied preprocessing methods on both datasets. To handle missing value in `Adult` dataset, we tried two different methods, namely delete instances with missing value and impute missing value. Then, we used One Hot Encoding to convert categorical attributes and we normalized numerical attributes. We then investigated the performance of K-nearest neighbours and decision trees on both datasets. Finally, we concluded that decision trees has overall better performance than K-nearest neighbours, and runs significantly faster. Besides, we evaluated the effect of changing method of handling missing value, changing hyperparatmeter, reducing the size of dataset on the performance of K-nearest neighbours and decision trees.

1 Introduction

The task of the project is explore two datasets. The first dataset is `Adult` dataset from UCI Machine Learning RepositoryDua:2019. The goal of `Adult` dataset is to predict whether an adult make over 50K a year [1]. It contains 14 attributes and 1 output. The second dataset is `Avila` dataset [2] from UCI Machine Learning Reporsitary [3]. De Stefano et al. [2] performed experiments on digital images of "Avila Bible" to prove that page layout features are helpful for writer identification. The features and classes extracted from the Avila Bible images is then made available to public by `Avila` dataset [2]. `Avila` dataset [2] contains 10 attributes and 12 output classes that represents 12 writers.

In order to to train models and make prediction based the two datasets, we firstly applied preprocessing methods on them. After loading them into `pandas` [4] dataframes, in order to handle missing value, we tried two different methods: 1) simply delete instances with missing value; 2) numerical normalization + impute missing value using Simpleimputer from datawig [1] (deep learning method), especially on categorical features, which serves as optional part.

In the experiments, we implemented cross-validation on two supervised learning methods: K-nearest neighbours and decision trees on `Adult` and `Avila` datasets. After cross-validation, we could get the evaluation metrics for two methods. We used `KNeighborsClassifier` and `DecisionTreeClassifier` of `scikit-learn` [5] to train and test datasets by K-nearest neighbours and decision trees. We investigated the impact of changing hyperparameters and reducing the size of dataset on both methods.

Since there is a large number of features after OneHot Encoding, in order to reduce the time required for training, especially for training of K-nearest neighbours, we applied Principal component analysis (PCA) to reduce the number of features. We implemented it by using `PCA` class of `scikit-learn` [5].

The result shows that for both datasets, decision trees have overall better performance than K-nearest neighbours. The hyper-parameters would result in changes in performance of two methods. When we change the K value in K-nearest neighbours or change the depth value in decision trees, their performance also change most of the time. When we reduced the size of

[1] `https://github.com/awslabs/datawig`

training dataset, the testing accuracy and validation accuracy decreased, especially when the size is becoming too small.

2 Datasets

index	age	workclass	fnlwgt	education	education-num	...	income
0	39	State-gov	77516	Bachelors	13	...	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	...	<=50K
2	38	Private	215646	HS-grad	9	...	<=50K
3	53	Private	234721	11th	7	...	<=50K

(a) Adult Dataset

index	intercolumnar distance	upper margin	lower margin	...	writer
0	0.266074	-0.165620	0.320980	...	A
1	0.130292	0.870736	-3.210528	...	A
2	-0.116585	0.069915	0.068476	...	A
3	0.031541	0.297600	-3.210528	...	A

(b) Avila Dataset

Table 1: Part of Adult and Avila datasets from UCI Machine Learning Repository

Table 1 (a) shows part of `Adult` dataset. It contains 14 attributes, 32561 instances of training data, 16281 instances of testing data. The output of the dataset is income – "<=50K" or ">50K". There are some missing value in the dataset. The missing values are all in workclass, occupation and native-country attributes.

Table 1 (b) shows part of `Avila` dataset. It contains 10 attributes, 10430 instances of training data, 10437 instances of testing data. The output of the dataset is writers represented by 12 classes – "A", "B", "C", "D", "E", "F", "G", "H", "I", "W", "X", "Y".

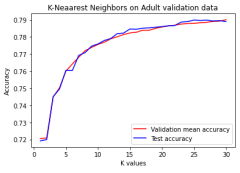
3 Results

For the performances comparison, decision trees has better overall performance than K-nearest neighbours. It could be obviously seen in the following more detailed tests.

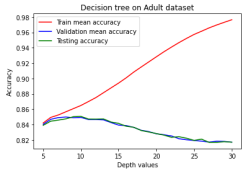
method	**K-nearest neighbours**	**decision trees**
delete instances	0.776	0.839
impute	0.788	0.814

Table 2: Accuracy when using different methods of handling missing values in Adult dataset

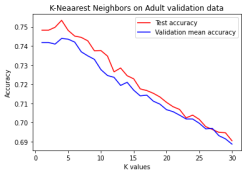
Table 2 shows the accuracy of K-nearest neighbours and decision trees when different methods in preprocessing is applied in `Adult` dataset. It shows that impute missing value has better result on K-nearest neighbours, while delete instances with missing value would result in better performance on decision trees. The reason is that K-nearest neighbours highly relies on previous training data to predict, so less instances of training data would result in bad performance.



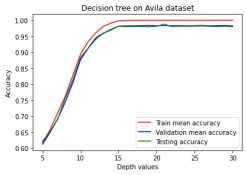
(a) KNN on Adult dataset



(b) decision trees on Adult dataset



(c) KNN on Avila dataset

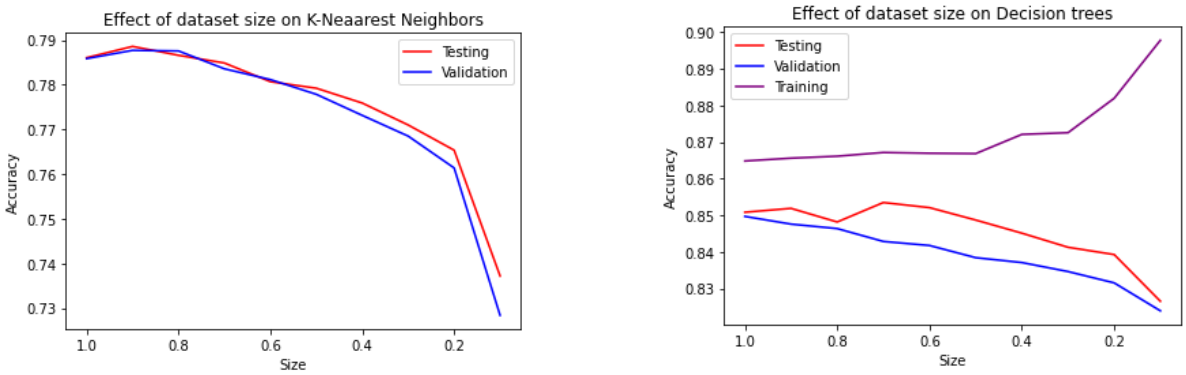


(d) decision trees on Avila dataset

Figure 1: Effect of changing hyperparameters for KNN and decision trees on two datasets

Fig. 1 shows the effect of changing hyperparameters. We changed K in K-nearest neighbours and depth in decision trees, then experiment them on both `Adult` dataset and `Avila` dataset. It could be seen that decision trees has overall better performance than K-nearest neighbours. For `Adult` dataset, lower K or higher depth value result in overfitting, which decrease the accuracy.

However, for `Avila` dataset, it is less likely to overfit because lower K or higher depth value still shows good performance within a certain range.



(a) K-nearest neighbours

(b) decision trees

Figure 2: Effect of changing dataset size for KNN and decision trees

Fig. 2 shows how the size of dataset impacts the accuracy for both methods. When we reduced the size of training dataset, the testing accuracy and validation accuracy decreased, especially when the size is becoming too small. The training accuracy keep increasing when the size is reduced, so it is clear that decreasing the size of dataset would result in overfitting.

4 Discussion and Conclusion

In this project, we explored `Adult` and `Avila` datasets, then use them to evaluate K-nearest neighbours and decision trees models. According to the result, for handling missing value in `Adult` dataset, impute missing value has better result on K-nearest neighbours, while delete instances with missing value has better performance on decision trees. For changing hyperparameters, changing K and depth has significantly different results on different dataset: K-nearest neighbours and decision trees are more likely to overfit on `Adult` dataset, on which decreasing K or increasing depth cause obvious drop in accuracy. However, they are much less likely to overfit on `Avila` dataset, on which decreasing K or increasing depth would cause a rise in accuracy within a certain range. According to experiments, decision trees has better overall performance and needs less time to train than K-nearest neighbours.

For further studies, more preprocessing techniques could be applied to datasets, and more datasets could be experimented to prove the performance comparison between K-nearest neighbours and decision trees.

5 Statement of Contributions

Saad Benslimane: Data preprocessing, apply K-nearest neighbours and decision trees
Yuhongze Zhou: Cross-validation, optional methods
Dun Yuan: Model evaluation, report writing

References

[1] Kohavi R, et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In: Kdd. vol. 96; 1996. p. 202–207.

[2] De Stefano C, Maniaci M, Fontanella F, di Freca AS. Reliable writer identification in medieval manuscripts through page layout features: The "Avila" Bible case. Engineering Applications of Artificial Intelligence. 2018;72:99–110.

[3] Dua D, Graff C. UCI Machine Learning Repository; 2017. Available from: `http://archive.ics.uci.edu/ml`.

[4] pandas development team T. pandas-dev/pandas: Pandas. Zenodo; 2020. Available from: `https://doi.org/10.5281/zenodo.3509134`.

[5] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.