Personality Analysis with Machine Learning in Spark

LI Xiang	XIA Yunlei	ZHANG Weiwen	ZHANG Yao
20492986	20583266	20583632	20583319

Abstract

This report aims to analyze a dataset about the Big Five personality traits. Some factor analysis can be applied on such personality surveys to explore the underline relationship between different aspects of the personality of the people. However, such analysis could be not enough if only a few samples are involved, and some big data and cloud computing technologies are involved to handle the huge amount of data. In conclusion, we have established several machine learning models like logistic regression and random forest to predict some values based on other survey results. Besides the prediction problems, we also applied statistical analysis like clustering to explore the data.

keywords: Apache Spark, logistic regression, decision tree, random forest, k-means

1. Dataset

The original dataset consists of 1,015,342 questionnaire answers. There are 50 questions in the survey, and the country, which is determined by technical information. All the 1M records are stored in a single CSV file. Each question is answered with a scale between 1 and 5. The scale was labeled as 1=Disagree, 3=Neutral, 5=Agree.

Since the dataset contains a huge number of records, without the help of cloud computing, both the data storage and time-consuming can be a big problem. To handle this challenge, we deployed Apache Spark and MLlib to finish the objectives.

The dataset can be accessed from https://www.kaggle.com/tunguz/big-five-personality-test

2. Data Pre-processing

Since the original dataset contains many null data, and the original data types didn't fit the requirement of MLlib. Data preprocessing is necessary.

1) Null filling: Due to the large dataset, it's costly and unnecessary to do imputation. We just removed records with null entries/
2) Remove zero: the questionnaire answers are on a scale of 1 to 5, however, many data in the dataset have zero value which means it is useless, so we will drop all the data entries with zero value. After this operation, there are 873173 entries left, about 20% of data were dropped.
3) One Hot for prediction: The original data is in the form of integers from 1 to 5. Since we want to predict the answer to the question as a discrete label, the continuous integers may not be suitable for the classification problem. We changed the label feature, which is randomly selected in each classification problem set, from the form of an integer to the form of a vector with 5 dimensions.

4) Data augmentation for clustering: Since the dataset is relevant to personality surveys, and the scores of different personalities can be calculated from answers. We add the calculated scores of different personalities after answers. These scores can be considered as dimension reduction and would be used in the clustering problem.

After the data preprocessing, the dataset contains 873173 records. As illustrated in figure 1, the values follow the normal distribution for different features.

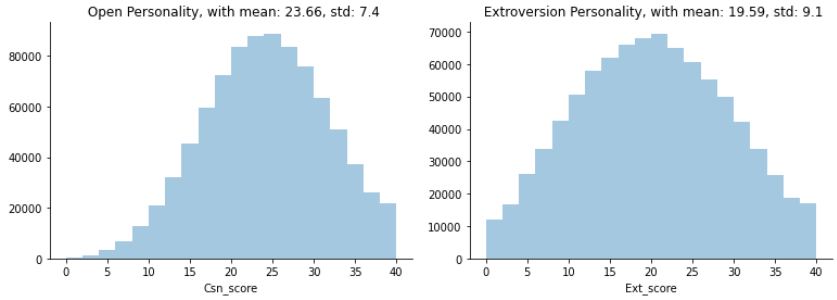

Figure 1: Answers to Different Questions

3. Our Approach

After data preprocessing, we randomly split the data into 80% and 20%, corresponding to the training set and test set. Apache Spark was performed on the Databricks community version. Instance with 16GB memory, 2 cores, and 1 DBU was deployed.

3.1 Spark MLlib and ML

MLlib is a strong scalable library in Spark for many classical machine learning algorithms. All executions in MLlib are working on RDDs. And ML is another powerful machine learning library in Spark but mainly working on Dataframes. In this project, logistic regression and random forest, decision trees are implemented with MLlib, and K-Means is implemented with ML.

3.2 Logistic Regression

Logistic regression essentially predicts the probability of a specific classification. Our task in this project is working on a 5-multiclass logistic regression to classify the first column “EXT1” with Mllib API.

3.3 Decision Tree

In pyspark.mllib.tree, we can use the DecisionTree API to implement decision trees. The max depth of the decision tree has a great influence on the accuracy of the result, so we checked the different results with the same max_bins but different max_depth. Their relationship is shown in figure 2.

According to figure 2, when the max_depth is smaller than 10, the accuracy of the decision tree also increases with the increasing of the max_depth. When the max_depth is around 9~11, the accuracy is around 48.3%. And when the max_depth is greater than 10, the greater of the max_depth, the lower of the accuracy. So, we choose max_depth = 10, and final accuracy reaches 48.673%.

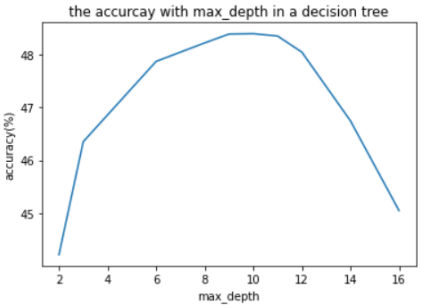

Figure 2: Relationship Between Max Depth and Accuracy

3.4 Random Forest

Random forests are an ensemble learning method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. Random forests have two important parameters which are max_depth and tree_num, we tested the influence of different values of these two parameters and got the following result. Their relationship is shown in figure 3 (left).

The accuracy will increase with the tree_num getting larger, but when the tree_num exceeds 40, the change of accuracy is not obvious. So, we choose the tree_num = 200 max_depth = 12, and the accuracy is 49.635%. Accuracies with different max_depth can be seen in Figure 3 (right).

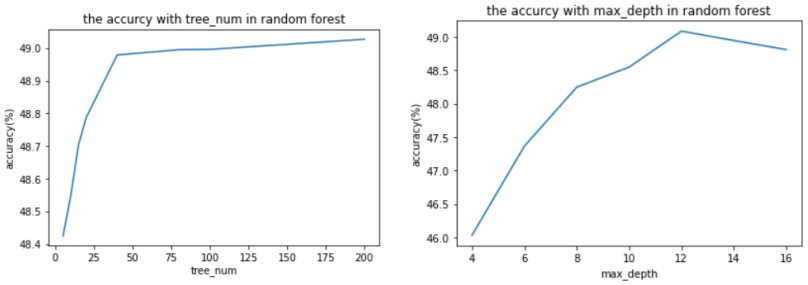

Figure 3: Relationship Between Parameters and Accuracy

3.5.K-Means

To better illustrate the data, the K-Means clustering approach is adopted. We tested the total scores for each personality based on [1] with 3 groups, 4 groups, and 5 groups. The visualized results are done with dimension reduction by PCA and are shown in Seaborn Figure 4. Based on these results of the centers, we may conclude that 4-clusters can describe the grouping situation of testers best. The centers for this clustering result are shown in table 1.

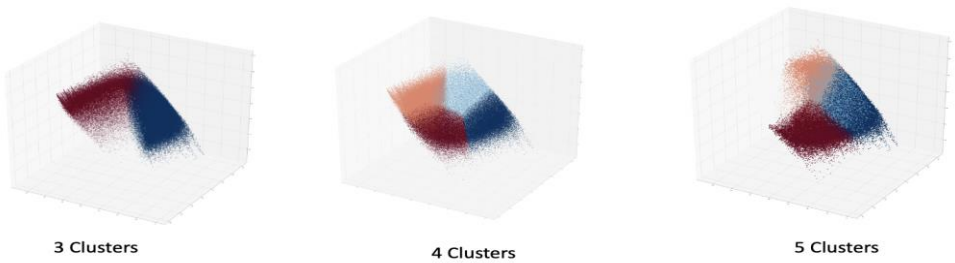

Figure 4: Visualized Results of Clustering

Ext_score	Est_score	Opn_score	Agr_score	Csn_score
14.061	24.568	28.019	23.733	26.045
28.749	27.879	31.193	26.146	27.633
25.977	14.481	29.516	25.357	20.697
10.749	11.240	27.547	23.836	20.693

Table 1: Locations of Centres for 4-Clusters

4. Result

Here we summarize the accuracies of our models in Table 2.

Model	Acc	Model	Acc
Logistic Regression	47.22%	Decision Tree	48.67%
Random Forest	49.64%	K-Means	N/A

Table 2: Accuracies of Different Models

Although the result is acceptable in consideration of the 5-classification problem, it is far away from state-of-the-art. Then we analyzed the correlation of the data in figure 5, and the result reveals that data within each one of the 5 categories show a strong linear correlation. The dependencies between variables cause the poor performance of machine learning.

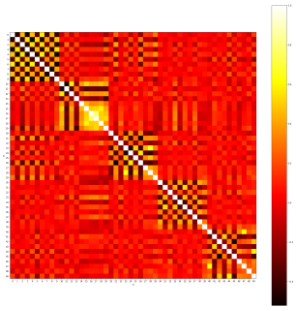

Figure 5: Visualization of Correlation Matrix

Also, we compared the importance of different features in one random forest model. As observed, only around 20% of features are important, others can be considered irrelevant. And even for these relevant features, the importance isn't very high, which means these features are not sufficient to predict the value.

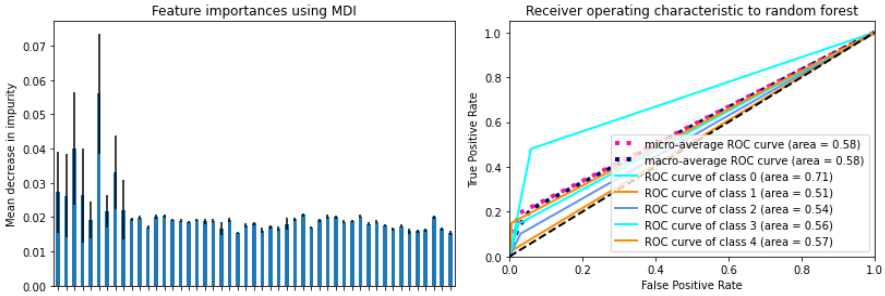

Figure 6: Importance of Features and ROC-AUC of Random Forest

After clustering, we could easily observe the five kinds of personality. To be noticed, since visualization of more than 3 dimensions is impossible, only 4 kinds are distinguishable in the figure 7. The bounds for different classes are clear, and we can easily classify records.

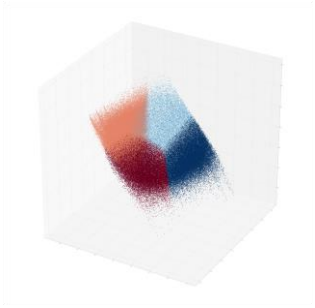

Figure 7: Clustering

Reference:

[1] “The Big Five Personality Test (BFPT).” .https://sites.temple.edu/rtassessment/files/2018/10/Table_BFPT.pdf