CPS844 Assignment 2

Xiaoxin Zhou 501072188
Chaoyu Wang 500815335

Background:

This report uses the "Car Evaluation Data Set" [1] to solve two practical problems: association analysis and clustering analysis. The structure of this report is followed by an introduction to the dataset, the methods we used, interpretation of the results and our conclusions. The data description from source link: https://archive.ics.uci.edu/ml/datasets/Car+Evaluation. Input attributes are printed in lowercase. Besides the target concept (CAR), the model includes three intermediate concepts: PRICE, TECH, COMFORT. Every concept is in the original model related to its lower-level descendants by a set of examples. The Car Evaluation Database contains examples with the structural information removed, i.e., directly relates CAR to the six input attributes: buying, maint, doors, persons, lug_boot, safety. Because of known underlying concept structure, this database may be particularly useful for testing constructive induction and structure discovery methods [1].

The Attribute information for the Car data set:

- **Buying:** v-high, high·, med, low
- **Maint:** v-high, high, med, low
- **Doors:** 2, 3, 4, 5-more
- **Persons:** 2, 4, more
- **Lug_boot:** small, med, big
- **Safety:** low, med, high

Class: unacc, acc, good, vgood

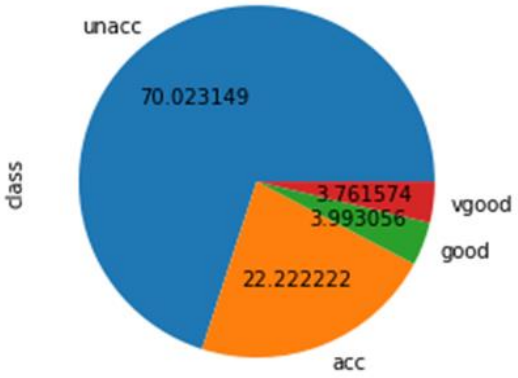

Figure 1: We visualized the data with a pie chart for class attributes, we found out the unacc class has a high proportion.

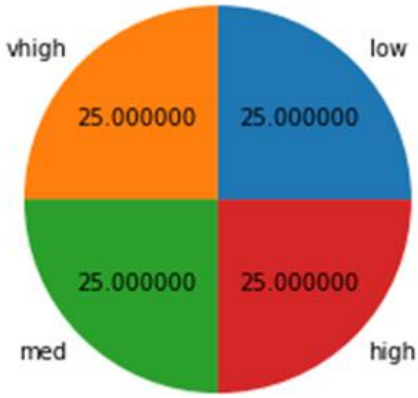

Figure 2: Visualize the data with a pie chart for buying and maint attributes are the same.

Figure 3: Visualize the data with a pie chart for doors attribute.

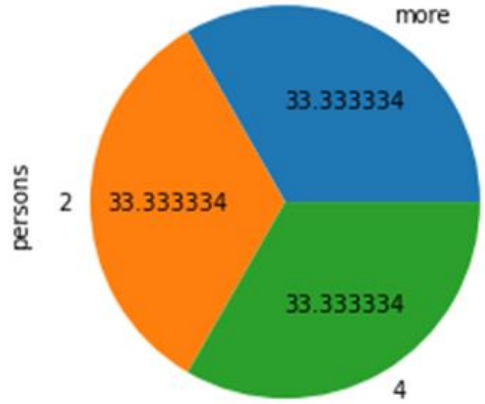

Figure 4: Visualize the data with a pie chart for persons attributes.

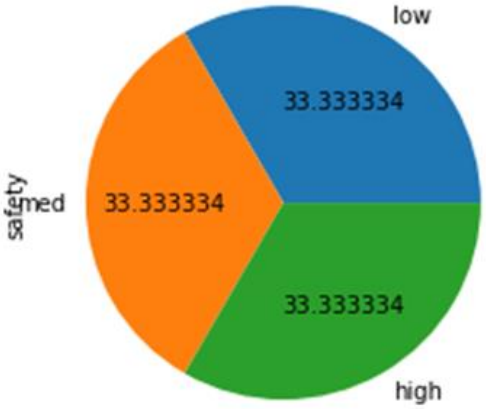

Figure 5: Visualize the data with a pie chart for safety attributes.

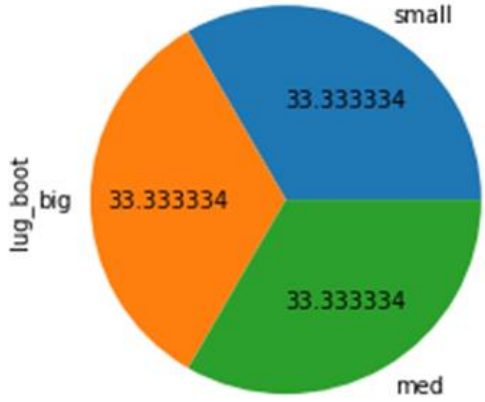

Figure 6: Visualize the data with a pie chart for the lug_boot attribute.

From the visualization, we found out the lug_boot, safety and persons are evenly distributed, and doors, buying and maint are also evenly distributed.

Method:

After we understand the dataset, we want to explain the method that we used in association analysis and clustering analysis.

Association Analysis: We call the function apriori to create the rules that meet the minimum support, confidence and lift requirements, then set the minimum support and confidence to 0.3 and 0.5. Then we want to print out each rule that was generated, along with its support and confidence.

Association analysis is discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. The example of the association analysis, {Soccer shoes} => (Soccer ball); when customers buy the soccer shoes, there is a high possibility for them to buy a soccer ball. That information can be used as the basis for decisions about marketing activities. [2] In our code, we are using Apriori to do the analysis. The explanation of Apriori: Apriori uses a "bottom up" approach, where frequent subsets are extended one item at a time, and groups of candidates are tested against the data. The algorithm terminates when no further successful extensions are found. Apriori uses breadth-first search and a Hash tree structure to count candidate itemsets efficiently. It generates candidate length itemsets from length itemsets. Then it

prunes the candidates which have an infrequent sub pattern. According to the downward closing lemma, the candidate set contains all frequent-length item sets. it scans the transaction database to determine frequent itemsets among candidates. [2].

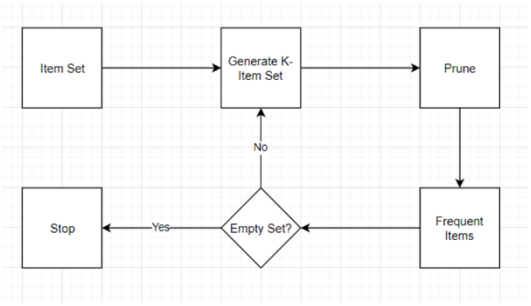

Figure 7: Association analysis structure.

Cluster analysis is a statistical method for processing data. It works by organizing items into groups or clusters based on how related they are. Unlike many other statistical methods, cluster analysis is often used when no assumptions about possible relationships in the data are made. It provides information about associations and patterns in the data, but not what these might be or what they mean. Clustering is measured using intracluster and intercluster distance. Intracluster distance is the distance between the data points inside the cluster. If there is a strong clustering effect present, this should be small (more homogenous). Intercluster distance is the distance between data points in different clusters. Where strong clustering exists, these should be large (more heterogeneous) [3]. In our code, we are using K-means. The K-Means algorithm determines the existence of clusters by finding their centroid points. The centroid point is the average of all the data points in the cluster. Each point can be assigned to a cluster by iteratively evaluating the Euclidean distance between each point in the dataset. The centroid points are random at first and change each time as the process progresses. K-means is often used in cluster analysis, but it has certain limitations, mainly for scalar data [3].

Result:

Association analysis:

We use different minimum support and confidence, initially we set the minimum support and confidence to 0.28 and 0.5 and print it. It can be found that the minimum support is not 0.28, but 0.3. So we reset the minimum support and confidence to 0.3 and 0.5. Through this adjustment, we got the following results and found that this is the minimum support and confidence we found.

[] --> ['2'] Support: 0.5 Confidence: 0.5
[] --> ['4'] Support: 0.5 Confidence: 0.5
[] --> ['high'] Support: 0.625 Confidence: 0.625
[] --> ['low'] Support: 0.625 Confidence: 0.625
[] --> ['med'] Support: 0.75 Confidence: 0.75
[] --> ['unacc'] Support: 0.7002314814814815 Confidence: 0.7002314814814815
['2'] --> ['high'] Support: 0.3125 Confidence: 0.625
['2'] --> ['low'] Support: 0.3125 Confidence: 0.625
['2'] --> ['med'] Support: 0.375 Confidence: 0.75
['2'] --> ['unacc'] Support: 0.4386574074074074 Confidence: 0.8773148148148148

['4'] --> ['high'] Support: 0.3125 Confidence: 0.625
['4'] --> ['low'] Support: 0.3125 Confidence: 0.625
['4'] --> ['med'] Support: 0.375 Confidence: 0.75
['4'] --> ['unacc'] Support: 0.30671296296296297 Confidence: 0.6134259259259259
['high'] --> ['low'] Support: 0.3333333333333333 Confidence: 0.5333333333333333
['high'] --> ['med'] Support: 0.4305555555555556 Confidence: 0.6888888888888889
['high'] --> ['unacc'] Support: 0.4085648148148148 Confidence: 0.6537037037037037
['low'] --> ['med'] Support: 0.4305555555555556 Confidence: 0.6888888888888889
['low'] --> ['unacc'] Support: 0.45601851851851855 Confidence: 0.7296296296296296
[] --> ['med', 'unacc'] Support: 0.5023148148148148 Confidence: 0.5023148148148148
['unacc'] --> ['vhigh'] Support: 0.3541666666666667 Confidence: 0.5057851239669422
['2'] --> ['med', 'unacc'] Support: 0.32407407407407407 Confidence: 0.6481481481481481
['med', 'low'] --> ['unacc'] Support: 0.30497685185185186 Confidence: 0.7083333333333333

Cluster analysis:

We passed a for loop to calculate the position of the best k, but because k-mean requires the input format to be different from our data. So we replace the data in the 6 labels with numbers, and calculate the SSE and numbers of cluster drawing. And use the elbow method of the elbow method to find the best k value. We found the best K for our dataset is 3. Then we set n_clusters equal to 3 and run it again. The silhouette score for the k-means classification is 0.181. We also get a scatter image to represent the cluster's model. But it can be seen that this is not ideal.

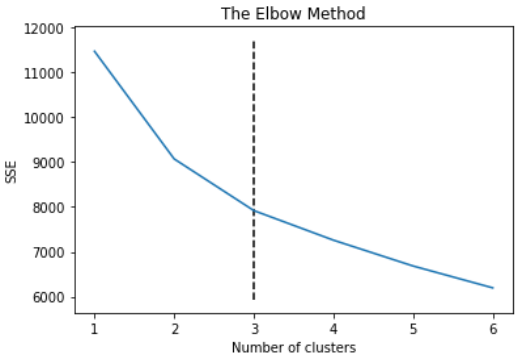

Figure 8: The Elbow Method result.

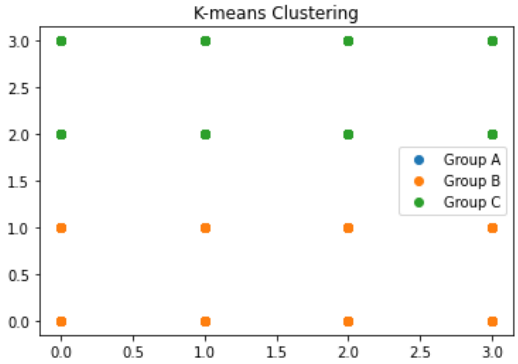

Figure 9: K-means Clustering result.

Conclusion:

Through the two analyses, it is found that the outputs obtained by the two analyses of clustering and association are completely different. The association method derives the relevant data and derives their minimum support and confidence. The cluster method has been classified more intuitively, but the association may not be as strong as the association method. But the cluster method of kmean is not suitable for our data, or it looks a bit strange because we did not make a 3d cluster diagram. But in

general, the two methods have their own different advantages, and they will have different effects in data mining.

Reference

[1] Marko Bohanec, “Car Evaluation Data Set”, UCL Machine Learning Repository, https://archive.ics.uci.edu/ml/datasets/Car+Evaluation

[2] Wikipedia, “Association rule learning”, Wikipedia, https://en.wikipedia.org/wiki/Association_rule_learning

[3] Qualtrics, “What is cluster analysis? When should you use it for your survey results?”, Qualtrics, https://www.qualtrics.com/experience-management/research/cluster-analysis/#:~:text=Cluster%20analysis%20is%20a%20statistical,how%20closely%20associated%20they%20are.