Clustering Analysis

Kilian Sennrich

20 4 2021

This is the report for the second assignment of the class "10616 Machine Learning". The task is to perform a clustering analysis on three problems provided by the lecturer. Time to complete task: 9 Days (Handin on April, 25., 2021)

Introduction of the Three Problems:

Problem 1 Based on the Switzerland_Population.csv dataset, which other canton is most similar to BS in terms of ...

a) age (3 features: Age0.19, Age20.64, Age64Plus)

b) religion (3 features: RomCath, Protestant, noConfession; other confessions = gap to 100%)

c) political parties (11 features: FDP, ..., OtherParties)

Problem 2 Perform a hierarchical cluster analysis on the Taxes_Cantons.csv dataset. Choose one method you consider most appropriate for this task. Pay special attention to pre-processing of data, the choice of distance measure; and the linkage method and argue your decisions.

Problem 3 The dataset mortgage.csv provides transformed and anonymised data of a subsample of a bank's mortgage customers. How many different types of customers are there? And how can these types be characterized? Again, choose one (at most two) method(s) you consider most appropriate for this task, and describe your reasoning.

Problem 1

Procedure for all Three Subquestions

At first the data was read in and checked for missing values (there were no missing values). All values of the relevant features represent percentages. The data was standardized during preprocessing, even though all features are on the same scale. Standardization is a controversial topic and it seems that there is not really a right or wrong. Some people recommend it only for features that are not on the same scale, others recommend it in any situation (full discussion: https://onlinelibrary.wiley.com/doi/10.1002/(SICI)1097-0266(199606)17:6<441::AID-SMJ819>3.0.CO;2-G). In my situation, the results for standardized and non-standardized features slightly differ. Since the data.table::fread() function, other than the read.csv(), does not allow to load matrices, the first row had to be converted to rownames. There were two approaches performed to derive the answer for the questions at hand. First, agglomerative nesting, a basic form of hierarchical clustering, was preformed and dendrograms were drawn (Figure 1). The average linkage method was used, since the dataset was rather small (and therefore not a computational problem). Average linkage

is also considered less prone to "wrongful" results than complete or single linkage (one might say the results are more robust). The closest canton to BS is the one that has the shortest distance on the connecting lines of the dendrogram. Second, the distances were directly evaluated via the distance matrix. The results will be exactly the same as with the hierarchical clustering approach (since hierarchical clustering relies on the distances from the distance matrix). The euclidean distance was used, since the features were all interval continuous in nature (percentages).

a) age (3 features: Age0.19, Age20.64, Age64Plus)

- The canton most similar to BS regarding age is **GR**. The standardized distance of the tho cantons is: 2.15.

b) religion (3 features: RomCath, Protestant, noConfession; other confessions = gap to 100%) The data consisted of three columns. One could think of adding a fourth column representing the other-Confessions feature, but this would not change the results, since this information is already covered by the other variables. Some people advise against in, since this can potentially lead to multicollinearity.

- The canton most similar to BS regarding Religion is **NE**. The standardized distance of the tho cantons is: 0.64.

c) political parties (11 features: FDP, . . . , OtherParties)

- The canton most similar to BS regarding political parties is **BL**. The standardized distance of the tho cantons is: 1.53.

Note: Possibly due to rounding some of the colums do not add up to 100%

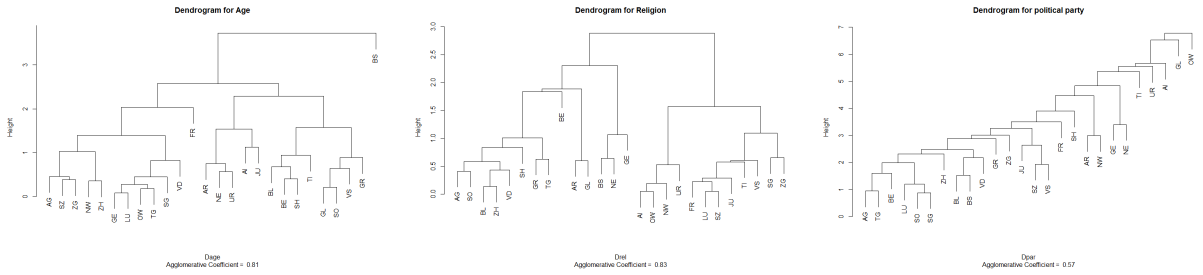

Figure 1: Left: Dendrogram for Age, Middle: Dendrogram for Religion, Right: Dendrogram for Political Party

Problem 2

Preprocessing In the data was read in and inspected for NAs. There were no NAs found, however, after a more indepth look into the data, it became obvious, that zeros mean NA. There is just no possibility that any canton has a zero value in property_transfer. Property transfer means the transfer of assets from the donor to heirs or donees. It is just plain impossible e.g. for canton Schwyz not to have any property transfers.

To get the best results from clustering, the data could just be replaced by values from the internet. However, since the year for this data is unknown, this is not really a practical approach. So the NAs had to be imputed. Multiple Imputation with chained equations was used for that matter. Since the data is very small, a good imputation can likely not be achieved. However, by deleting the columns or rows with NAs,

much information would be lost. To validate the MICE imputation, for each of the three columns including NAs, one value was converted to an NA. After that, MICE algorithm was run using *predictive mean matching* (PMM), which is a regression method (https://www.jstor.org/stable/1391390?origin=crossref&seq=1). In a paper explaining the MICE function, Stef van Buuren and Karin Groothuis-Oudshoorn write the following on the PMM method (https://www.jstatsoft.org/article/view/v045i03): *"Its main virtues are that imputations are restricted to the observed values and that it can preserve non-linear relations even if the structural part of the imputation model is wrong."*

It has to be noted, that this imputation method on this dataset is not likely to be very accurate. However, since there is not really a better way to deal with the data, it is still used. Since the "land" feature had almost 50% of missing values, recovering them with MICE was pretty much impossible. This is why this feature was not included in the final cluster. After the Imputation, the data was scaled, so that all variables contribute equally to the clusters.

Model Selection Since all features represent amounts of money, which follows an interval scale, the euclidean distance measure was used. For clustering, the bottom up approach "agglomerative nesting" was used. This approach suits in particular, when the different cantons should be put into clusters by their closest similarity. The average linkage method was used, since it is considered to be more robust than single or complete linkage (https://www.wiwiweb.de/spss-lernen/clusteranalyse/hierarchische-klassifikation/linkage-methoden.html) Since the data is rather small, computational expenses do not have to be considered for this approach.

Results The results are pretty easy to interpret. All the cantons that have large cities in them, namely, Zürich (Zürich), Waadt (Lausanne), Genf (Genf) and Basel Stadt (Basel), are dissimilar to the other cantons in terms of their taxes. Out of these 4 cities, especially Zürich stands out and forms its own cluster (Figure 2).

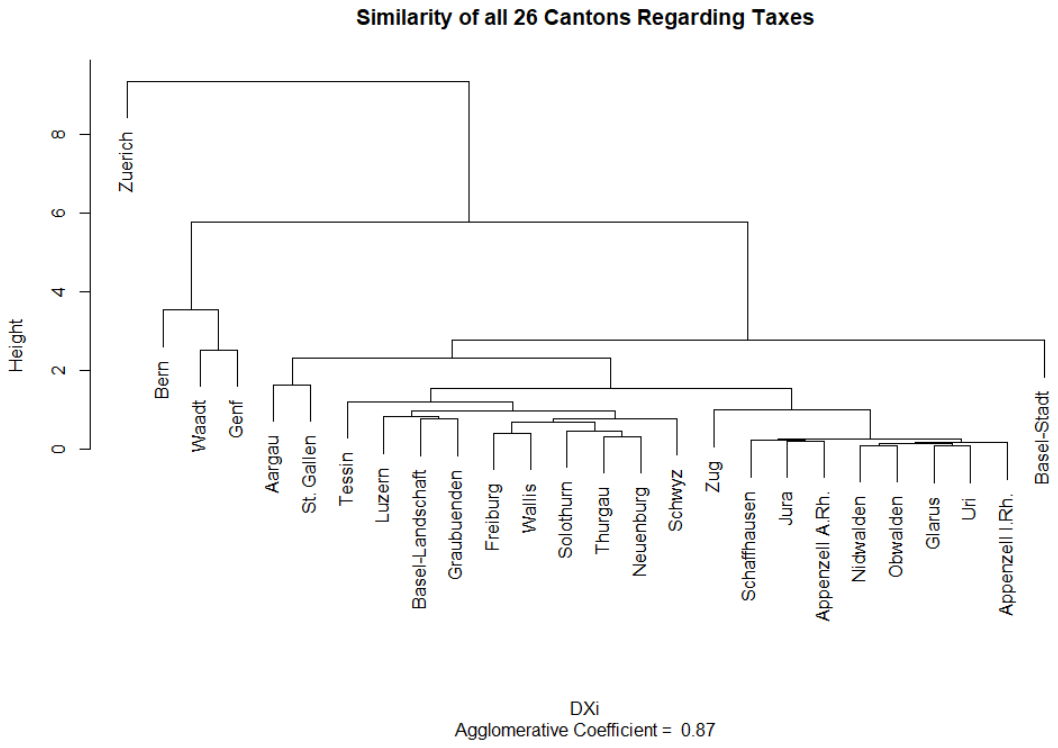

Figure 2: Dendrogram for Taxes

Problem 3

The dataset was read in and inspected. It was in a tidy format and no NAs could be indentified. To break down the dataset for one plot, a principal component analysis and multidimensional scaling were performed. The dataset got scaled at the outset.

As a clustering approach, k-medoid (Position Around Medoid: PAM) clustering was used. Since the data from the MDS-plot showed pretty widespread groups, it was assumed that there might potentially be outliers in the dataset. This was also checked with boxplots (most outliers lie in the annual income feature). Besides the outliers, it might be crucial to have observations that represent the "prototype" of the group (since it was specifically asked for it in the outlining of the problem: *And how can these types be characterized?*). PAM can better deal with these two problems/properties than k-means: PAM is rather robust to outliers and the medoid is always an actual observation. The downsides are however, that the amount of clusters needed are rather hard to determine, that redundant features don't get averaged out (other than in k-means) and that it is computationally expensive. The latter two turned out not to be a problem (the dataset was rather small (comp. power was not an issue), and had only three features anyways). Defining the number of clusters was a hard task however. By eye it is hard to determine whether two or three clusters will be needed (in various forums they don't recommend looking at the data but rather focus on metrics, however i think it can sometimes still be helpful). To select the final amount of clusters, I compared the silhouettes of two, three and four clusters (Figure 3). For the k = 2 silhouette, the average silhouette coefficient was the lowest (Figure 3 left) and there were some outliers in the second cluster. With k = 3, clusters 2 and 3 get much higher silhouette widths than with k = 2. Cluster 1 in oppose gets a much smaller silhouette width. The overall/average silhouette width increases. With k = 4 the group that already had a small silhouette width gets further split up into two clusters. How to interpret that? One of the "assumptions", in order for PAM to work properly, is not met for the dataset at hand: *Covariance homogenity* (A better way to deal with this would be *model-based clustering*, however, when applied to the data, the results look rather strange). One cluster is spread far further around the group medoid than the others. Therefore, the silhouette for this cluster is rather small. In the end I decided for three clusters, since the cluster, that is widespread around the medoid might be the reason for the low silhouette widths for k = 2. For k = 4 the widespread cluster gets divided into two subclusters but both subclusters have pretty small silhouette widths. One might say that one subcluster includes the "outliers" (that originated rather by the wider spread around the medoid) from the other cluster. Therefore 3 clusters were chosen in the end. In order to make the strength of the group memberships visible, additionally a fuzzy clustering approach was used.

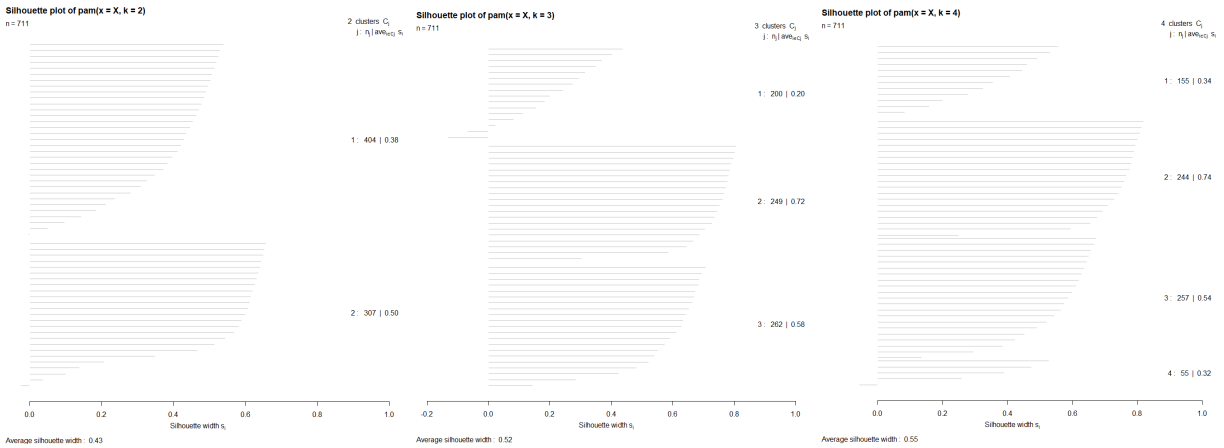

Figure 3: Left: k = 2, Middle: k = 3, Right: k = 4

Interpretation of the clusters The great strength of PAM is that it returns actual observations, that are the "prototype" of the cluster. The three medoids can be seen in table 1 and in Figure 4 below. The standardized medoids can be interpreted as follows: values close to zero are close to the mean of the feature,

negative values are below average, positive values are above average. It seems like there is a group of people (black), that have a good salary (above average), that do not save much money and have a mortgage that is slightly lower than the average mortgage. These might be people that are willing to spend the money they earned. It could be that some of them bought a big house (went for the highest mortgages they could possibly get) and thus can save less. The large covariance in the group indicates the heterogeneity of the the people that are willing to spend. The members of the red group have an equivalent income to the black group but have saved much more money. They have also payed of their mortgage to a further extent. It might be that people from this group have not gone for the maximum mortgage they could get and have therefore payed off much of their mortgage already. They could therefore also save more money. The green group might consist of below average earners, that could only slowly pay of their mortgage and that could save a normal amount of money. Overall it can be sayed that there are the below average earners (green) and the above average earners. The above average earners can be further divided into people that are willing to spend their money (black) and people that are willing to save their money (red). The people that are willing to spend their money are rather heterogeneous.

Table 1: standardized values for the medoids

Medoid	annual_income	savings	mortgage
black	0.2465849	-1.307864	-0.3069275
red	0.2145848	1.090886	-0.9590822
green	-0.8617813	-0.078054	1.0717544

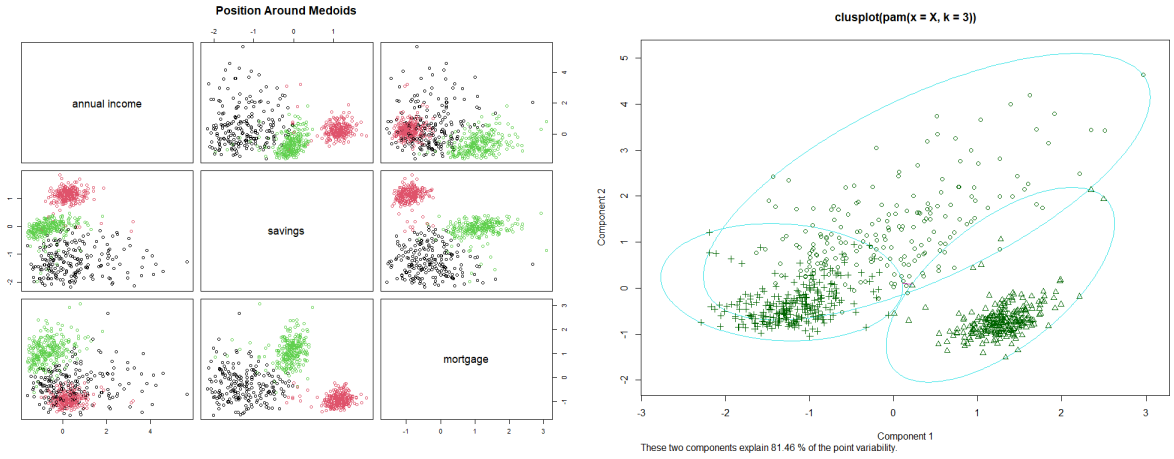

Figure 4: Left: Clustering of the three groups, Middle: Right:condensed plot

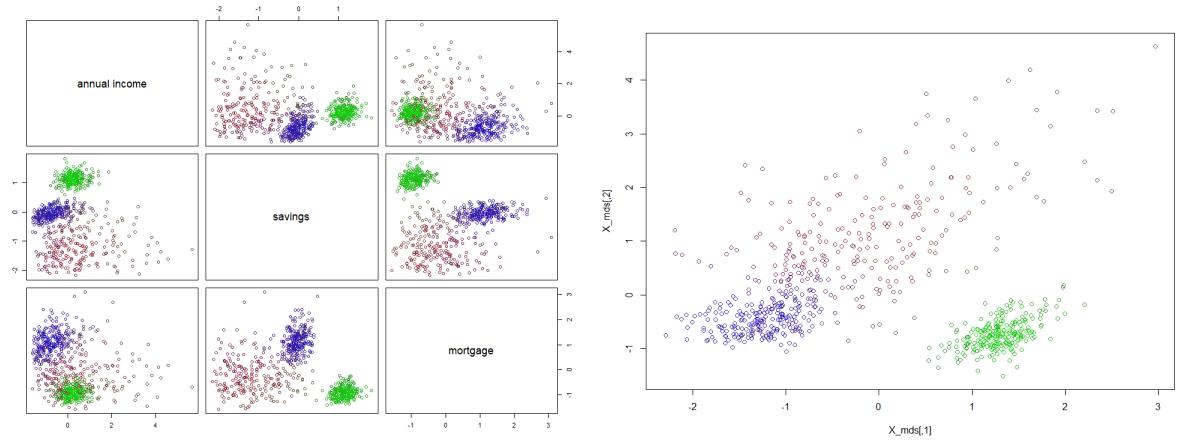

Figure 5: Plots indicating the strength of the group membership, (Observations might be classified differently than with PAM); Here the red group is the black group of Figure 4; the violet group is the green group and the green group is the red group.