Cluster Analysis

Prepared by Amee Amin

Assignment 8

MSPA Course 410, Summer 2017

Professors Chad R. Bhatti and Tom Miller

Introduction

At the heart of most research agendas is a drive to understand hidden patterns of human behavior, and leverage them for business insights, biological innovation, and improvements in any industry. Research studies that follow how people manage their time can be helpful for companies ranging from lifestyle or home improvement products, health centers trying to understand how lifestyles impact health outcomes, or even for companies who need to better understand a potential market for their product. In this analysis we compare two forms of clustering analysis of a time-use dataset, consisting of 28 observations of 10 time-use variables and 5 demographic factors from a time-use survey study conducted across many countries in the mid-1970s.

In previous exploratory data analysis, we used the dimensionality reduction techniques of PCA and Factor Analysis to reveal hidden underlying factors among the time-use variables, including a strong association between professional, transport, housework, and childcare. Leisure explained the most variance in our data. In this analysis, we explored K-means and hierarchical clustering as alternative methods of understanding hidden patterns among time-use activities and demographic factors in the time-use dataset, and we expected to similar clustering among professional, transport, housework, and childcare in this analysis as well. We found that the outcome of K-means clustering for both time-use and demographic variables had a larger silhouette coefficient than hierarchical clustering, and we recommend management use K-means clustering to understand consumer groups for future target marketing.

Section 1: Review of Prior Research

Section 1.1: Dataset Overview

Our time-use dataset contained 10 time-use variables and 5 demographic factors with 28 observation groups. Demographic data includes group, gender, professional work status, marital status, and country. The dataset was created for the purpose of explaining and testing classification techniques. The time-use variables describe general categories of activities carried out in daily life: professional (time spent working), transport (time spent in transportation), housework (time spent doing housework), childcare (time spent taking care of children), shopping (time spent shopping), personal (time spent doing personal activities), mealtime (time spent eating), sleep (time spend sleeping), tv (time spent watching tv), and leisure (time spend doing activities for leisure). The time-use variables reflect the average number of minutes spent in a fixed-time observation period of 40 hours. We also noted that exercise, a common daily activity, was not included in this dataset. The average amount of time spent on activities ranged from 0 to approximately 850 minutes. The demographic variables include group, gender, professional work status, marital status, and country.

Out of 28 observations, 16 were missing professional work status and 12 were missing marital status. It appears as though two or more datasets were merged together to create the time-use dataset. The missing data may make it difficult to get conclusive insights from the clustering of demographic variables.

Table 1.1: Correlation Matrix of Time-Use Variables

	professional	transport	housework	childcare	shopping	personal	mealtime	sleep	tv	leisure
professional	1.000000	0.939146	-0.906398	-0.864779	-0.654015	-0.112085	-0.461478	-0.558446	-0.055886	-0.250538
transport	0.939146	1.000000	-0.870412	-0.810035	-0.503057	-0.077350	-0.610170	-0.704684	-0.041169	-0.164435
housework	-0.906398	-0.870412	1.000000	0.861274	0.499720	-0.039976	0.358374	0.437793	-0.205751	-0.053089
childcare	-0.864779	-0.810035	0.861274	1.000000	0.541772	0.117891	0.364119	0.280785	0.121612	-0.054544
shopping	-0.654015	-0.503057	0.499720	0.541772	1.000000	0.590629	-0.182912	-0.021966	0.218571	0.229963
personal	-0.112085	-0.077350	-0.039976	0.117891	0.590629	1.000000	-0.353476	-0.211088	0.324885	0.038271
mealtime	-0.461478	-0.610170	0.358374	0.364119	-0.182912	-0.353476	1.000000	0.818196	0.317928	0.065580
sleep	-0.558446	-0.704684	0.437793	0.280785	-0.021966	-0.211088	0.818196	1.000000	0.019653	0.272753
tv	-0.055886	-0.041169	-0.205751	0.121612	0.218571	0.324885	0.317928	0.019653	1.000000	-0.071656
leisure	-0.250538	-0.164435	-0.053089	-0.054544	0.229963	0.038271	0.065580	0.272753	-0.071656	1.000000

We looked at a correlation matrix of the 10 time-use variables to quickly explore the relationships between variables. Professional and transport appear to be strongly positively correlated, while professional and housework appear to be strongly negatively correlated. Housework and transport are therefore also strongly negatively correlated as well. Childcare expectedly has a strongly negative correlation with professional and transport and a strongly positive correlation with housework. Shopping has a negative correlation with professional and transport, and a positive correlation with housework and childcare. Personal time has a negative correlation with professional, transport, and housework, while it has a positive correlation with childcare and shopping. Mealtime negatively correlates with professional, transport, shopping, and personal, while it positive correlates with housework and childcare. Sleep negatively correlates with professional, transport, shopping and personal, and positively correlates with housework, childcare, and mealtime. Leisure, tv and sleep have weak correlations with other variables.

Section 1.2: PCA and Factor Analysis Overview

We applied PCA to our explanatory variables using the sklearn package. The principle components are the dimensions along which our data is most differentiated, as expressed by one, two, or more variables. In layman's terms, the principal components reflect an underlying dimension that summarizes or accounts for the original variables.

After performing a PCA, the first component, which explained almost half of the variance in our data, was most strongly associated with leisure in the negative direction. This is consistent with our exploratory data analysis that showed leisure had the weakest correlations with other variables, and therefore explained the most variance. The second component, which captured approximately 0.2 of the variance, was tv in the negative direction. The third component, which captured approximately 0.13 of the variance, was most strongly associated with sleep in the positive direction. The fourth component, which captured approximately 0.12 of the variance, was the most strongly associated with professional in the positive direction. The fifth component, which captured approximately 0.04 of the variance, was most strongly associated

with personal in the positive direction. The sixth component, which captured approximately 0.02 of the variance, was most strongly associated with shopping in the positive direction.

Comparatively, we performed a Factor Analysis to determine the "common underlying factors" that maximize shared variance, instead of total variance. In factor analysis, you assume that the latent constructs are the actual causal factors underlying the collinearity among your variables. Therefore, factor analysis looks not only at a covariance matrix of variables but also a covariance matrix of noise.

The results of the factor analysis showed that the variances for professional, sleep, and leisure were fully explained by the factors. The variances for transport, housework, childcare, shopping, mealtime, and tv were almost fully explained by the factors (at least 0.90 of their variance). Last, only 0.628 of the variance for personal was explained by the factors. Overall, the factors capture a majority of the variance explained by the time-use variables.

We recommended that the Factor Analysis rotated solution be used (instead of PCA) by management because it is easier to interpret how each factor represents a different latent relationship among initial explanatory variables.

Section 2: Distance Measures and Input Matrices

Preprocessing data is important for clustering analysis because clustering relies upon calculating distance. Using the sci-kit learn package, we pre-processed the time-use variables to center the features around zero with unit variance. We then used numpy to transpose the preprocessed matrix for clustering analysis (example in Figure 2.1 below).

Figure 2.1 Transport Matrix Data for Professional (Time-Use Variable)

```
array([[  7.22982085e-01,   1.17292306e-01,  -1.96897249e+00,
          7.45415040e-01,  -1.21073862e+00,   6.10817312e-01,
          1.48698443e-01,   9.11418906e-01,   2.74322990e-01,
         -1.92410658e+00,   9.24878678e-01,  -1.26009112e+00,
          8.66552996e-01,  -2.68554516e-01,   9.02445724e-01,
          4.98652538e-01,  -1.96897249e+00,   9.02445724e-01,
         -8.47324749e-01,   7.45415040e-01,  -7.11445137e-02,
          9.02445724e-01,   5.79411175e-01,  -1.90616021e+00,
          9.11418906e-01,  -6.66579228e-02,   7.99254132e-01,
         -7.11445137e-02],
```

Because the demographic variables are categorical, we transformed each demographic variable into dummy variables. Using dummy variables, we inputted a total of 38 demographic variables into the clustering analysis after transposing the data. We did not pre-process the data, as the demographic data was not scalar.

In the next section's clustering analysis, we specified a Euclidean distance because K-means assigns points to the nearest centroid using Euclidean distance from each point to a centroid, as opposed to pairwise distances between data points.

Section 3: Clusters of Activities

We performed both a K-means clustering analysis and Hierarchical clustering analysis to obtain clusters of time-use variables. The silhouette coefficient from the hierarchical clustering analysis with four clusters was 0.33. We specified an agglomerate, or "bottom up" strategy for hierarchical clustering, where each observation starts in its own cluster, and pairs of clusters and joined as the process moves up a hierarchy. We chose four clusters after observing four clusters with similar branch length in a dendrogram plot. The silhouette coefficient from the K-means clustering analysis with six clusters was 0.38. In this analysis we will interpret the results from the K-means clustering analysis.

With no preconceived notions about the number of clusters, we searched across 2 through 9 cluster analysis solutions and evaluated the silhouette coefficient for each. The silhouette coefficient reveals how similar an object is to its own cluster versus other clusters. Ranging from -1 to 1, where a high value signifies that an object matches its own cluster well but not neighboring clusters, the coefficient takes both cohesion within a cluster and separation among clusters into account.

Figure 3.1 Results of K-Means Analysis Solutions for n-Clusters

```
nclusters:  2  silhouette coefficient:  0.365424558428
nclusters:  3  silhouette coefficient:  0.314388336183
nclusters:  4  silhouette coefficient:  0.333350753678
nclusters:  5  silhouette coefficient:  0.377370568187
nclusters:  6  silhouette coefficient:  0.380656762221
nclusters:  7  silhouette coefficient:  0.337416451292
nclusters:  8  silhouette coefficient:  0.244135708094
nclusters:  9  silhouette coefficient:  0.151817122951
```

N=9 clusters performs the most poorly with a coefficient of 0.15, and N-8 clusters also performs poorly with a coefficient of 0.24. No substantial structure was found for n=9 or n=8 clusters. N=6 clusters has the strongest coefficient of 0.38, which indicates that a weak and potentially artificial structure was found. N=5 clusters and N=2 clusters have the next highest coefficients of 0.37, followed by n=7 clusters, n=4 clusters, and n=3 clusters, which all similarly imply a weak or artificial structure.

Assuming there are 6 natural clusters of time-use variables, given that it had the highest silhouette coefficient, we evaluated the clustering solution attributes for a K-means clustering of n=6.

Figure 3.2 K-Means Clustering Solution of n=6

```
   cluster  variable
4        0  shopping
5        0  personal

   cluster      variable
0        1  professional
1        1     transport

   cluster variable
8        2       tv

   cluster  variable
6        3  mealtime
7        3     sleep

   cluster   variable
2        4  housework
3        4  childcare

   cluster variable
9        5  leisure
Silhouette coefficient for the five-cluster k-means solution:  0.380656762221
```

The six clusters reflect the results from our previous studies. The variable leisure is in its own cluster, and it previously explained most of the variance in our PCA analysis. The variable tv is also in its own category, and it previously explained the second-most variance in our PCA analysis. The variables of professional and transport are in their own cluster, and they had an r-squared value of 0.94 (Table 1.1). The variables of housework and childcare are in their own cluster, and they had an r-squared value of 0.86 (Table 1.1). The variables of sleep and mealtime are in their own cluster, and they had an r-squared value of 0.82 (Table 1.1). Last, shopping and personal were also in their own cluster, with an r-squared value of 0.59 (Table 1.1).

The clusters group together activities that would be grouped together in daily life. Tv may be described as a form of leisure, shopping a form of personal activity, and childcare a form of housework. Moreover, transportation is needed for most forms of professional work. Mealtime and sleep, interestingly, are two categories of “essential” human needs — in other words, everyone needs to eat and sleep as bodily functions, so it is not surprising they are in a similar cluster. However, the highest silhouette coefficient of 0.38 is still low. Given the collinearity of time-use variables and small dataset size, we concluded that a weak or artificial clustering structure may be very likely.

Section 4: Clusters of Demographic Groups

We performed both a K-means clustering analysis and Hierarchical clustering analysis to obtain clusters of demographic variables. The silhouette coefficient from the hierarchical clustering analysis with four clusters was 0.31. We specified an agglomerate strategy for hierarchical clustering, and we chose nine clusters after observing 9 clusters with similar branch lengths in a

dendrogram plot. The silhouette coefficient from the K-means clustering analysis with six clusters was 0.47 In this analysis we will interpret the results from the K-means clustering analysis.

With no preconceived notions about the number of clusters, we searched across 2 through 37 cluster analysis solutions and evaluated the silhouette coefficient for each. A portion of the results are shown below.

Figure 3.1 Results of K-Means Analysis Solutions for n-Clusters

```
nclusters:  2  silhouette coefficient:  0.47248697792
nclusters:  3  silhouette coefficient:  0.382626706071
nclusters:  4  silhouette coefficient:  0.341026502425
nclusters:  5  silhouette coefficient:  0.322643439605
nclusters:  6  silhouette coefficient:  0.321445949465
nclusters:  7  silhouette coefficient:  0.314610806409
nclusters:  8  silhouette coefficient:  0.311990298092
nclusters:  9  silhouette coefficient:  0.310333520006
nclusters:  10  silhouette coefficient:  0.310702934371
nclusters:  11  silhouette coefficient:  0.251449497565
```

N=2 clusters performs the most strongly with a coefficient of 0.47, which indicates a weak or artificial structure, but is close to values that resemble a reasonable structure (a silhouette coefficient between 0.51 and 0.70). Clusters from n=3 through n=10 have silhouette coefficients in the range of 0.30-0.40, which indicates that the clustering has a weak or artificial structure. N=11 clusters has a silhouette coefficient below 0.25, implying that no substantial clustering structure was found.

N=12 through N=31 clusters had negative silhouette coefficients between 0 and -3.5, which implies we should consider clustering above N=11 with caution. N=32 through N=37 clusters all had a silhouette coefficient of 0, which we should also consider with caution. Low or negative vales may imply too few or too many clusters.

Assuming there are 2 natural clusters of demographic variables, given that it had the highest silhouette coefficient, we evaluated the clustering solution attributes for a K-means clustering of n=2. The results show that one cluster consists of the gender “women”, and the other cluster consists of all the other variables. One reason for this imbalance could be that there are nearly double the amount of men in the dataset than women. We recommend further studies explore binarizing the categorical variables, as opposed to creating dummy variables. Researchers should also test imputing or removing the missing values.

Section 5: Model Comparison and Recommendation

Based on the results from this analysis, we would recommend the K-means clustering model for understanding consumer groups for future target marketing. This is due to one main reason: given that there is not a large amount of natural clustering in the data, hierarchical clustering

may not be needed in terms of computational power. Hierarchical clustering takes far more time than k-means clustering, especially as the dataset size increases.

More broadly, the results of our analysis do not show any clear consensus on which methods produces better clustering, per se, because the silhouette coefficients were all below 0.50. Leisure and tv seem to consistently appear different from the other variables in this dataset, and their ability to distinguish customers should be explored further. In general, hierarchical clustering can be useful in exploratory analysis when you do not know the number of clusters that need to be specified. Moreover, hierarchical clustering produces a repeatable outcome, whereas K-means begins with a random choice of cluster centers. We recommend that hierarchical clustering and dendograms be used as supplemental information on the natural clustering of time-use and demographic variables at this point. Further research on the natural clustering of time-use and demographic variables is needed before management invests in hierarchical clustering solutions.

Lastly, we recommend that the findings in this analysis be taken with caution. The dataset size is very small and missing many missing demographic observations. Future studies should seek to use a larger dataset and use alternative methods of data imputation or cleaning.

Conclusion

We recommend a K-means clustering model for management to understand consumer groups for future target marketing. We found both K-means clustering and hierarchical clustering as helpful methods of revealing natural clustering, or the lack thereof, in our data. As previous studies showed, we found indication that leisure and tv explain much of the difference in our data and show their own, individual natural clustering away from the other variables. Professional and transport continue to show clustering together, as well as housework and childcare. K-means clustering of demographic variables resulted in women in their own cluster, compared to all the other demographic variables in their own cluster. We recommend investing in securing better demographic data for future analyses, such as methods to impute missing data or gathering a larger dataset with a higher proportion of women. A larger dataset could help illuminate more characteristics of the natural clustering. If, for example, the natural clusters were non-spherical, then hierarchical clustering may be a better method. However, a more robust dataset is needed to pursue any conclusive analysis.