SYDDANSK UNIVERSITET

Clustering

Bjarki Sigurdsson
Abdulrahman Abdulrahim
Miguel de la Colina

April 4, 2017

ABSTRACT

THE PURPOSE OF THE EXERCISE WAS TO USE CLUSTERING TO SEE HOW THIS WOULD AFFECT THE PERFORMANCE ON THE KNN CLASSIFICATION BUT NOW TAKING INTO CONSIDERATION THE CENTROID OF THE CLUSTERS ALSO CREATING DENDOGRAMS TO GET A VISUAL REPRESENTATION OF HOW THIS CLUSTERS ARE MADE AND RUN EVALUATION METHODS TO SEE HOW WELL THEY ARE PERFORMING.

1 KMEANS CLUSTERING

For this part of the exercise we will try to improve the performance by performing kmeans in each cipher individually for the training set in order to represent them as cluster centroids and then perform knn using the centroids we have obtained.

We will be training the kmeans algorithm using different sizes of clusters (25, 50, 100, 200) to see how the accuracy will change in bot the person-dependent and person-independent set.

As we can see from the results that we received from table 1 we can see that the accuracy level actually increases as we increase clusters which makes sense seeing that as we begin creating more clusters we are being more specific to finding to which cluster they belong giving us a more accurate result than doing it with a fewer number of clusters.

Table 1: Accuracy Results.

clusters	p-dependent accuracy	p-independent accuracy
25	.83875	.14025
50	.87	.143
100	.87675	.149
200	.89275	.17225

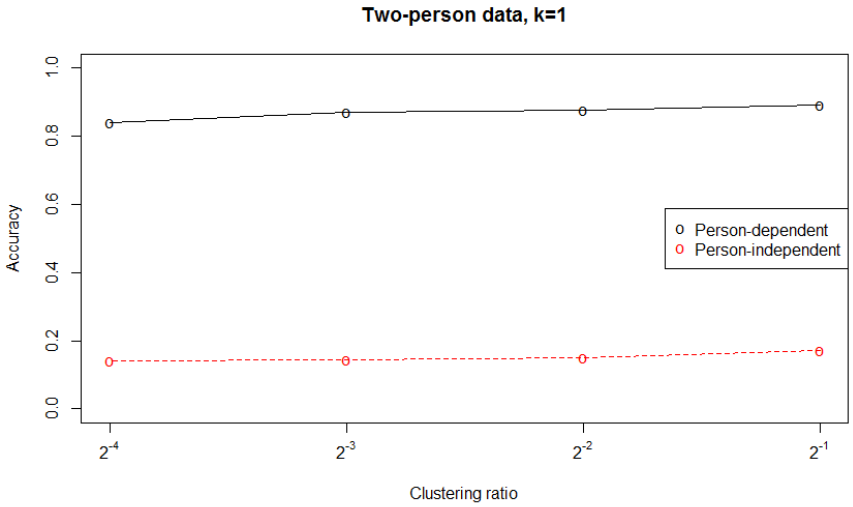

Figure 1: Accuracy graph for dependent and independent data-sets.

In figure one we can see how the accuracy varies for each of the data sets depending on how many clusters we are using and here it is easier to see how the accuracy does increase even if it's a little while we increment the number of clusters.

2 HIERARCHICAL CLUSTERING

A dendogram is a tree diagram that will show how the clusters have been made so that we can see the relationships we have with each of the ciphers here we will see how this relationships look and how they relate with the cross validation tables from knn.

If we look at figure 4 and figure 5 we can see that there is a relation between both of them we can se that there are some clusters that actually are matched with different numbers at lower levels like with 7 and 1, and if we look at 7 and 1 in the cross validation table we can actually see that 1 and 7 have a number of misplaced results where the 1 is mistaken for a 7 which actually explains why we could be seeing this types of clusters at lower levels.

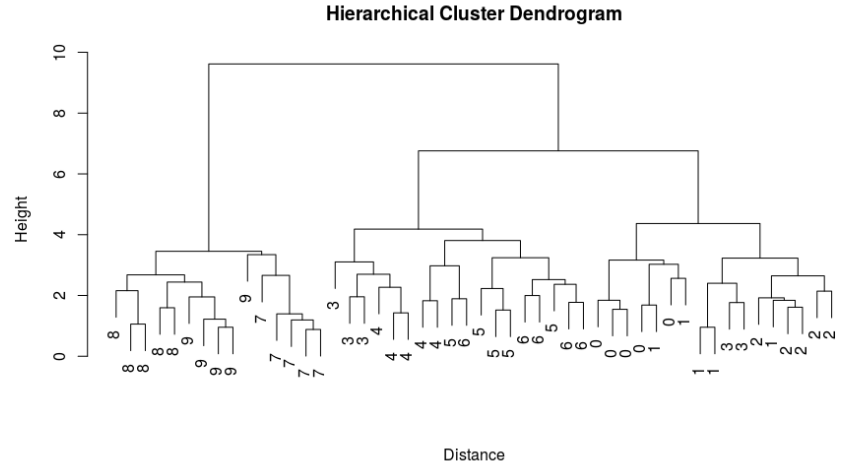

Figure 2: Low level dendogram containing 5 instances of each digit.

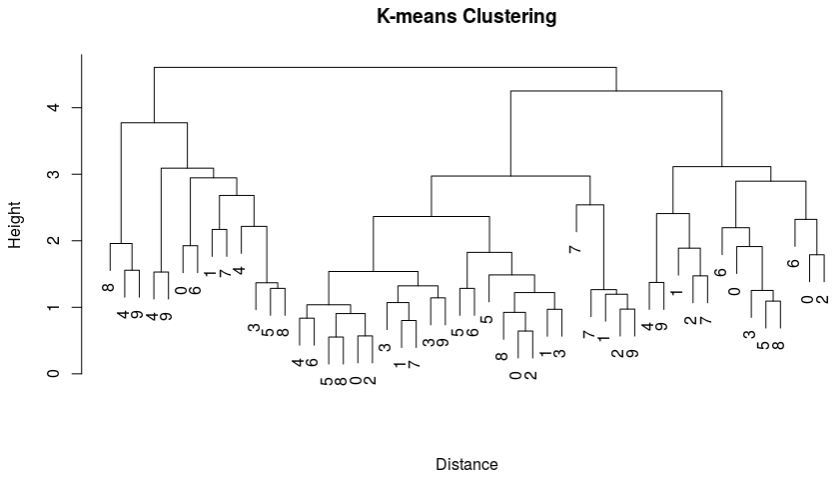

Figure 3: kmeans clustering.

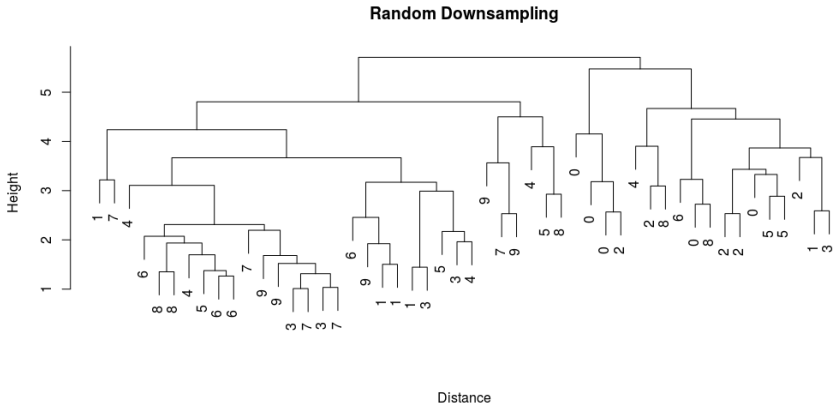

Figure 4: Random down sampling.

	0	1	2	3	4	5	6	7	8	9
0	289	14	21	12	24	16	8	3	11	15
1	9	257	12	24	7	3	22	38	4	17
2	47	43	265	79	14	6	23	24	33	21
3	1	36	20	215	11	32	13	15	22	13
4	4	7	1	6	278	10	9	0	5	25
5	27	7	11	19	15	263	14	7	55	7
6	12	8	5	5	12	7	263	0	12	0
7	3	25	18	26	6	9	3	249	2	17
8	5	2	20	17	26	64	17	3	262	1
9	4	1	19	7	17	10	15	57	4	258

Figure 5: cross validation table knn.

3 Evaluation methods

As we can see in figure 6 the precision recall curve shows that if we have low recall we will be having a high level of precision and as the recall begins to increase we will start to see how the precision decreases and although with each k the curve varies we can still see that all of them follow similar behaviors.

Now in figure 7 we have a graph which displays the F1-score (which takes into account both the precision and the recall) against the number of clusters we have decided to use and as we can see the F1-score begins to decrease noticeably as the number of clusters starts to increase.

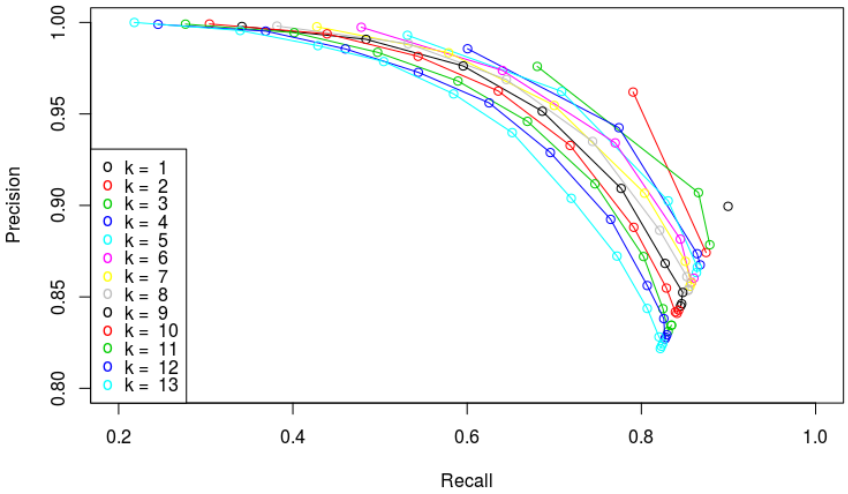

Figure 6: Precision-Recall curve.

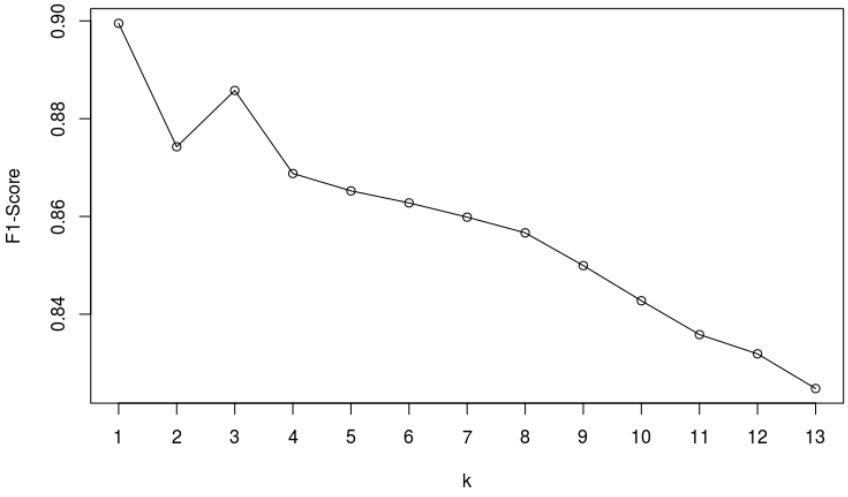

Figure 7: F1 score vs k graph.