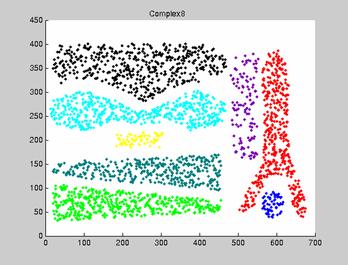
# Dr. Eick

Assignment3 COSC 3337 Fall 2019

Similarity Assessment, Clustering and Using Clustering to Create Background Knowledge for Classification Tasks

Individual Project[[1]](#footnote-1)

Third Draft

Due: Sunday, December 1, **11a** (for students who did not chose the late option for Assignment1 submissions will be accepted until Monday, December 2, 11p the latest!)

Last modified: November 20, 9a

Tentative weight of Assignment2: about 33-45% of the points allocated to the three assignments

**Learning Objectives**:

1. Learn to use popular clustering algorithms, namely K-means, K-medoids/PAM and DBSCAN
2. Learn how to summarize and interpret clustering results
3. Learn to write R functions which operate on the top of clustering algorithms and clustering results
4. Learning how to make sense of unsupervised data mining results
5. Learn how clustering can be used to create useful background knowledge and classification problems.
6. Learn how to create distance function and distance matrices in R and learn about their importance for clustering.
7. You will learn to devise search procedures to solve “*finding a needle in a hay stack problems*” that are typical for data mining projects.

**Datasets**: In this project we will use the *Complex8 dataset[[2]](#footnote-2)—depicted above— which is a 2D dataset; it can be downloaded at:* <http://www2.cs.uh.edu/~ml_kdd/restored/Complex&Diamond/Complex8.data>; you also find some visualizations of this dataset at:[*http://www2.cs.uh.edu/~ml\_kdd/restored*](http://www2.cs.uh.edu/~ml_kdd/restored)and the “cleaned” Pima Indian dataset (the original can be found at: <https://archive.ics.uci.edu/ml/datasets/pima+indians+diabetes>)which is a “null-value cleaned” version of the original dataset, called Pima dataset, in the following; it can be found at: <http://www2.cs.uh.edu/~ceick/UDM/Pima.csv>

The Pima dataset has 8 numerical attributes and a binary class variable (1 indicates that the person is assumed to have diabetes), indicating the following information:

1. Number of times pregnant   
2. Plasma glucose concentration a 2 hours in an oral glucose tolerance test   
3. Diastolic blood pressure (mm Hg)   
4. Triceps skin fold thickness (mm)   
5. 2-Hour serum insulin (mu U/ml)   
6. Body mass index (weight in kg/(height in m)^2)   
7. Diabetes pedigree function   
8. Age (years)   
9. Class variable (0 or 1)

5 Examples in the modified Pima Indians Diabetes Dataset:

6,148,72,35,156,33.6,0.627,50,1

1,85,66,29,156,26.6,0.351,31,0

8,183,64,29,156,23.3,0.672,32,1

1,89,66,23,94,28.1,0.167,21,0

0,137,40,35,168,43.1,2.288,33,1

**Assignment3 Tasks**:

**0.** Transform the Pima dataset into a dataset ZPima by z-scoring the first 8 attributes of the dataset, and copying the 9th attribute of the dataset \*

**1.** Write an R-function entropy(a,b) that computes the entropy and the pecentage of outliers of a clustering result based on an apriori given set of class lables, where *a* gives the assignment of objects in O to clusters, and *b* contains the class labels of the examples in O. The entropy function *H* is defined as follows:

Assume we have m classes in our clustering problem; for each cluster Ci we have proportions pi=(pi1,…,pim) of examples belonging to the m different classes (for cluster numbers i=1,..,k); the entropy of a cluster Ci is computed as follows:

H(pi)= Σj=1 (pijlog2(1/pij)) (H is called the **entropy function**)

Moreover, if pij=0, pijlog2(1/pij) is defined to be 0.

The entropy of a clustering X is the size-weighted sum of the entropies on the individual clusters:

H(X)= Σr=1 (|Cr|/|Σp|Cp|)H(pr)

In the above formulas ”|…|” represents the set cadinality function[[3]](#footnote-3).

Moreover, we assume that X={C1,…,Ck} is a clustering with k clusters C1,…,Ck,

You can assume that cluster 0 contains all the outliers, and clusters 1,2,…,k represent “true” clusters; therefore, you should ignore cluster 0 and its instances when computing H(X). The entropy function returns a vector: (<entropy>,<percentage\_of\_outliers); e.g. if the function returns (0.11, 0.2) this would indicate that the entropy is 0.11, but 20% of the objects in dataset O have been classified as outliers. Run you function for the two test cases, listed at the end of the document, and report the results in your report. \*\*\*

**2.** Write an R-function wabs-dist(u,v,w) that takes two vectors u, v as an input and computes the distance between u and v as the sum of the absolute weighted differences—using the weights in w[[4]](#footnote-4).\*

**3.** Write an R-function create-dm(x,w) that returns a distance matrix for the objects in dataframe x by calling the wabs-dist(a,b,w) for all pairs of objects a and b belonging to x. \*\*

**4**. Run K-means on the ZPima dataset for k=6 and k=9 and nstart=20; next run PAM for k=6 with distance matrices that have been created using the following weight vectors for the ZPima dataset using the function create-dm─obtaining 3 different PAM clustering results:

a. (1,1,1,1,1,1,1,1)

b. (0.2,1,0,0,0,1,0.2,1)

c. (0,1,0,0,0,1,0,0)

For the 5 obtained clustering results report the overall entropy, the entropy of each cluster, the majority class and the centroid/medoid of each cluster. Next, visualize the clustering result of the K-means run for k=6/9 and for the 3 PAM results in the Plasma glucose/Body mass index (Attribute 2&6 Space) for the original dataset[[5]](#footnote-5). Interpret the obtained results! Does changing the distance metric affect the PAM results? Do the results tell you anything which attributes are important for diagnosing diabetes, and about the difficulty diagnosing diabetes? \*\*\*\*\*\*\*\*

**6.** Run DBSCAN on the ZPima dataset; try to find values for MinPoints and epsilon parameters, such that the number of outliers is less than 15% and the number of clusters is between 2 and 12[[6]](#footnote-6); visualize the obtained clustering result in the Plasma glucose/Body mass index attribute Space on the original dataset and report its entropy. Comment on the quality of the obtained clustering result. \*\*\*

**7.** Run K-means with nstart=20 for k=8 and k=11 for the Complex8 dataset; visualize the results, compute their entropy, and discuss the obtained clustering results and its quality.\*\*

**8.** Next, run DBSCAN for the Complex8 dataset; try to find “good” parameter settings[[7]](#footnote-7) for epsilon and Minpoints such that entropy of the obtained clustering result is minimized and the number of outliers is less than 8%[[8]](#footnote-8). Develop a search procedure looks for the “best” DBSCAN clustering for the Complex8 dataset by varying the epsilon/Minpoint parameters. Report the best clustering result you found, report the epsilon and Minpoint parameters you used to obtain this clustering result and report its percentage of outliers and entropy. Results with higher entropy will receive higher scores; students who find the “best” result will get extra credit worse \*\*’s! Also compare the DBSCAN results with those you obtained for K-means in Task 7, and assess which clustering algorithm did a better job in clustering the Complex8 dataset! \*\*\*\*\*\*\*\*

**9**. Summarize to which extend the K-Means, PAM, and DBSCAN where able to rediscover the classes in the Complex8 and Pima/ZPima datasets! Moreover, did your expermental results reveal anything interesting about diabetes? \*\*

**Training Cases for R-Functions that need to be developed in task1:**

*1st case*

Let a, and b be the following vectors:

a=(0,1,1,1,1,2,2,3)

b=(A,A,A,E,E,D,D,C)

entropy(a,b)=4/7\*H(0.5,0,0,0,0.5)+ 2/7\*0 + 1/7\*0=4/7  
  
2nd case

a=(1,1,1,0,0,2,2,2)

b=(A,A,A,E,E,D,D,C))

entropy(a,b)= ½\*0+1/2\*H(0,0,1/3,2/3,0)=1.37/3=0.456

**Training Cases for R-Functions that need to be developed for tasks 2 and 3:**

Let x be a data frame containing vectors a, b and c where:

a: (1, 2)

b: (4, 5)

c: (9, 12)

w is weight vector: (0.2, 0.3)

wabs-dist(a,b,w) = 1.5

wabs-dist(b,c,w) = 3.1

create-dm(x,w)=

|  |  |  |
| --- | --- | --- |
| 0 | 1.5 | 4.6 |
| 1.5 | 0 | 3.1 |
| 4.6 | 3.1 | 0 |

**Note:** The \* at the end of each problem denotes its approximate weight. Therefore, the higher the number of \*’s, the more points that problem will carry.

**Submission Guidelines:**

Create a folder and name it as LastName\_StudentId\_HW3. HW3 folder should include:

* R code for the tasks.
* The data files needed to run the R codes.
* The assignment report containing all the plots and results along with the interpretations for them.

Submit the LastName\_StudentId\_HW3 folder in a zipped file through Blackboard.

**Points will be deducted for incomplete submission**

1. No collaboration with your class mates is allowed! [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)
3. E.g. |{2,3,5}|=3 and |∅|=0! [↑](#footnote-ref-3)
4. It basically, computes sum((abs(a-b))\*w) [↑](#footnote-ref-4)
5. Add the $cluster column to the Pima dataset (and not the ZPima dataset) and visualize the clusters. [↑](#footnote-ref-5)
6. The reported result does not need to be optimal with respect to the entropy of the obtained clustering! [↑](#footnote-ref-6)
7. You might write a program that runs DBSCAN automatically for different parameter settings; you might also change the distance function, using the same procedure we used for PAM, to obtain “better” results. [↑](#footnote-ref-7)
8. That is, if you report a clustering result with more than 8% outliers, you will get a low score for this task! [↑](#footnote-ref-8)