Uppsala University
Department of Information Technology
Kjell Orsborn

DATA MINING I - 1DL360

Assignment 2 - k-Means and DBSCAN clustering

1 k-Means and DBSCAN clustering

In this assignment we focus on two different clustering techniques, namely k-Means clustering and DBSCAN.

The k-Means clustering method is a partitional clustering method. It is one of the most commonly used clustering methods as it is quite easy to understand and implement. It applies a proximity measure for forming the clustering.

DBSCAN [1] is a density-based clustering method that applies a density measure for forming the clustering.

In this assignment you will study both algorithms and examine their characteristics on two different 4-dimensional data sets. In this case, the data sets are completely synthetically generated. The reason for this is that they include specific hidden characteristics to be revealed by you in the assignment.

You will use an implementation of k-Means and DBSCAN in the AmosMiner application using the Amos II database management system. This means that you have to install Amos II and AmosMiner according to the instructions on the assignments home page.

2 Preparation

We suggest that you read about k-Means and DBSCAN in Chapter 7 in Tan et al. [Tan06]. As for Assignment 1, it also useful to study the sections on normalization and proximity measures that are treated in Chapter 2 in Tan et al.

You can find and download the AmosMiner and Amos II system from the assignments home page and you will also get a chance to familiarize yourself with those systems in two introductory tutorials.

There will also be an introductory tutorial to this assignment, which is not mandatory, but advisable to attend.

The computer labs include a limited number of computers and, to avoid overcrowded labs, each group should sign up for one lab time.

Lab lists will be posted on the board outside 1346.

3 Assignment

The data sets for this assignment reside in the following two files, which are accessible from the lab course homepage:

`clusterdata1.nt` including 150 rows that each include 4 fields.

`clusterdata2.nt` including 200 rows that each include 4 fields.

For both files this means that each row represents a data object through 4 attribute values.

You should perform cluster analysis on both these data sets using the k-Means and DBSCAN clustering algorithms implemented within the AmosMiner system. Furthermore, you should experiment with the different parameters used in those algorithms to see how they influence the clusterings. The different steps in the assignment are described in more detail below. There is also a result form (available on the assignments home page) to be filled in by you and handed in to your assistant. Also

see Section 4 on more details on examination and how you should report the assignment.

Thus, in your assignment you should do the following:

1. START-UP

   Once you have installed the AmosMiner application, you should have the script file and data files for assignment 2 available (if some file is missing in your AmosMiner directory you should download and install AmosMiner again).

   Load the script file for this assignment, `a2.osql`, in AmosMiner by the following command:

   ```
   < 'a2.osql';
   ```

   When the script file has been loaded correctly, you have already been executing a k-Means and a DBSCAN analysis that can be tracked by studying the content of the script file itself and the results that were generated during its execution. It is now your turn to modify these scripts by following the rest of these instructions.

2. NUMBER OF CLUSTERS

   You should experiment with different numbers of clusters to be identified with k-Means. Use different techniques:

   - projecting original data to different 2- and 3-dimensional subspaces
   - projecting to eigenvectors (Principal Component Analysis), as shown in STEP 1 in the `a2.osql` script file.

3. NORMALIZATION

   You should experiment with different normalization methods, such as:

   - Gaussian: $\bar{x} = \frac{x - mean(x)}{stdev(x)}$

- uniform: $\bar{x} = \frac{x - min(x)}{max(x) - min(x)}$

Also note that values of each dimension are normalized separately.

Hint: It is useful to use the built-in aggregate functions: *mean*, *standard deviation* and *maximum* and *minimum* object in a bag of objects respectively (see Amos II manual).

```
average(Bag of Number x) -> Real
stdev(Bag of Number x) -> Real
maxagg(Bag of Object x) -> Object
minagg(Bag of Object x) -> Object
```

This is an example of normalization on identified data using min-max normalization.

```
set :subtract = aggv(values_of_known_data(), #'minagg');
set :divide = aggv(values_of_known_data(), #'maxagg')
              - :subtract;

normalize_iddata(#'known_data', :subtract, :divide,
    #'normalized_known_data');
```

4. INITIAL CENTROIDS

   (a) Determine the most suitable initial centroids for each dataset. This can be one of random samples, or vectors of values produced by any other means (including visual analysis of scatter plots).

   (b) Save the "best" initial centroids you encounter to 'vic1.nt' for the data set in 'clusterdata1.nt' and 'vic2.nt' for the data set in 'clusterdata2.nt', as shown under STEP 3 in the `a2.osql` script file.

5. VISUALIZATION

   Experiment with different projections when plotting clusters in 3D. Determine most suitable projection for each dataset.

6. DETERMINE OPTIMAL Eps VALUE FOR DBSCAN

   (a) First, set the `minpts` value to 5. Create a graph of the 5-dist value of the data points and use this to estimate the amount of noise in each data set.

   (b) Then make a choice of `eps` that gives you the correct amount of noise.

   (c) Run DBSCAN algorithm for each data set and calculate SSE.

7. COMPARING THE ALGORITHMS

   (a) Which algorithm have you found to be most suitable for clustering of each dataset?

   (b) Which parameters (including initial centroids) did you use?

   You should motivate your answers and explain your decisions and reasoning.

You can also find these instructions and hints in the TODO list of the `a2.osql` script file.

You are also referred to the Amos II manual for further reading, which is available through the lab course home page. We suggest that you read section 2.6 about Collections. You might find the following sections especially useful: `average`, `stdev`, and `count` in 2.6.1, `groupby` in 2.6.2, and `aggv`.

4 Examination

At the examination, your implementations will be executed in Amos II by running the script files as follows:

```
< 'a2.osql';
```

When the execution is done, we expect you to present the outcome of your analysis when it comes to choice of different parameter values including:

- number of clusters, i.e. parameter `k` (KMEANS).
- set of initial centroids saved in for instance to the files `vic1.nt` and `vic2.nt` (KMEANS).
- `eps` parameter (DBSCAN).
- `minpts` parameter (DBSCAN).

Be prepared to answer questions regarding topics such as:

- A brief description of how the data was preprocessed and why.
- Which normalization, if any, results in better clustering (determine this visually, you cannot compare SSE values)? Why?
- How did you choose your `k` parameter in KMEANS?
- How will the selection of centroids influence the solution in KMEANS?
- How did you choose your `eps` and `minpts` parameters in (DBSCAN) and how might this choice influence your results?
- If it is known a priori (beforehand) that each attribute has different importance, how can that knowledge be used in the pre-processing to improve the classication performance?
- The Euclidean distance measure was used here. Do you think that the current measure is reasonable for the current data sets and can you think of other situations where it might be less appropriate?

References

[EK96] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, pages 226-231, 1996 (The paper is available on the lab course homepage).

[Tan06] Tan, P-N, Steinbach, M. and Kumar, V.: Introduction to Data Mining, Addison-Wesley, 2006.