# Theory Assignment 02

**Course Code: CSE-475**
**Course Title: Data Mining**

**Student Information**

| | |
|---|---|
| ID | 17182103062 |
| Name | Wasek Samin |
| Intake | 38 |
| Section | 02 |
| Shift | Day |

17182103062
Wasek Samin

10.12)

Present conditions under which density-based clustering is more suitable than partitioning based clustering and hierarchical clustering. Give application examples to support your argument.

Ans: Density based clustering is more suitable if no or very limited domain knowledge is available to determine the appropriate values of the parameters, the clusters are expect of arbitrary shape including non-convex shapes, and the efficiency is essential on large data sets. For example, consider the application of recognizing residential area on a map where buildings and their types are labeled. A user may not have the domain knowledge how a residential area would look like. Moreover, a residential area are may be of arbitrary shape, and may not be in convex, such as those built along a river. In a large city, there are hundreds of thousands of buildings. Efficiency on large data sets is important.

10.13)

Give an example of how specific clustering methods can be integrated, for example, where one clustering algorithm is used as a preprocessing step for another. In addition, provide reasons as to why the integration of two methods may sometimes lead to improve clustering quality and efficiency?

Ans: Consider building an ontology of queries for information reterioral and web search. A query is associated with a vector of web pages that are clicked when the query is asked. An effective approach is to first conduct density based clustering to form small clusters as micro-clusters Each micro-cluster is treated as a base unit, sometimes called a concept. Then, a hierarchical clustering method can be applied to build an ontology. To use density based clustering as a preprocessing step can quickly merge synonyms, such as "University of Illinois at Urbana-Champaign" and "UIUC", into one concept. This preprocessing step can reduce the data so that the ontology built later. is more meaningful, and the ontology building process is faster.

10.14)

clustering is recognized as an important data mining task with broad applications. Give one application example for each of the following cases:

a) An application that uses clustering as a major data mining function.

Ans: There are many uses of data clustering analysis such as image ~~prep~~ processing, data analysis, pattern recognition and ~~may~~ many more. Clustering in data mining helps in the classification of animals and plants are done using similar functions on genes in the field of biology. It helps in gaining structure of the species. Clustering is also used in outlier detection applications such as detection of credit card fraud.

Clustering helps in understanding each cluster and its characteristics. One can understand how the data is distributed and it works as a tool in the function of data mining

b) An application that uses clustering as a preprocessing tool for data preparation for other data mining tasks.

Ans: Web search engines often provide query suggestion services. When a user inputs one or a series of queries the search engine tries to suggest ~~on~~ some queries that may capture the user's information need. To overcome the data sparcity that is, many queries are asked very few times, and many web pages are not clicked. Clustering is often used as a preprocessing step to ~~o~~ obtain micro-clusters of queries, which represent similar user intents and web pages, which are ~~alot~~ about the same topics.

10.16)

Describe each of the following clustering algorithms in terms of the following criteria : 1) shapes of clusters that can be determined; 2) input parameters that must be specified; and 3) limitations.
a) k-means, b) k-medoids, c) ~~Clar~~ CLARA, d) BIRCH, e) CHAMELEON, f) DBSCAN

Ans: K-means :

1) Compact clouds (clusters of non-convex shape cannot be determined.

2) Number of clusters.

3) Sensitive to noise and outliers, works

good on small datasets only.

b) K-medoids:

1) Compact clouds (clusters of non-convex shape cannot be determined).

2) Number of clusters.

3) Small datasets (non scalable).

c) CLARA:

1) Convex-shaped clusters.

2) Number of clusters.

3) Sensitive to the selection of initial samples.

d) BIRCH:

1) Spherical in shape clusters.

2) N d-dimensional data points.

3) Resulting clusters may be of unnatural shape.

e) CHAMELEON:

1) Arbitrary shape.

2) N d-dimensional categorical points

3) Quadratic time in the worst case.

f) DBSCAN:

1) Arbitrary shape.

2) Maximum possible distance for a point to be considered density reachable and minimum number of points in a cluster.

3) Quadratic time in the worst case.