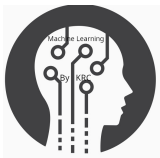
Machine Learning (Clustering Techniques)

Prof K R Chowdhary

CSE Dept., MBM University

December 28, 2024



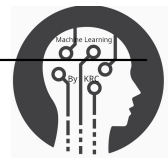
Introduction Clustering

⇒ The process of clustering partitions a set of data, according to some similarity measure, into several groups such that "similar" records are in the same group, so that each group represents a similar subpopulations in the data.

⇒ As an example, each cluster could be a group of customers, which have similar purchase histories or interactions or some other factors or the combinations [1].

Table 1: Data groupings of similar objects

Cluster No.	<Qty, Unit Price>
Cluster 1	<2, 1800>
	<3, 2050>
	<5, 2270>
Cluster 2	<15, 1800>
	<18, 2200>
	<12, 2380>
Cluster 3	<3, 250>
	<4, 180>
	<4, 200>



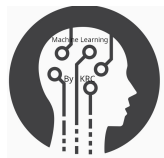
Introduction Clustering...

⇒ Clustering is based on common properties of items in each group/cluster, as follows: customers in cluster 1 purchased few high-priced items, customers in cluster 2 purchased many high-priced items, and customers in cluster 3 purchased few low priced items.

Definition

Cluster feature(CF). Collective summarized representation of a cluster to optimize space as well as to facilitate faster access. □

⇒ CF is a triple: cluster centroid, cluster radius, and number of points in the cluster.
⇒ CF based approach is efficient due: 1. they consume less space as all objects in a cluster are not required, and 2. they constitute sufficient information for computing all intra-cluster and inter-cluster. So distances can be computed fast.



Introduction Clustering...

$\Rightarrow$ Some points in clusters can be discarded, while the others can be compressed, as defined below.

Definition

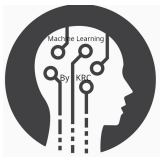
Discardable Point. A point is considered discardable, if its membership can be ascertained with high confidence. □

Definition

Compressible Point. A point that is not discardable, but belongs to a tight subcluster consisting of a set of points that always move between clusters simultaneously, is called a compressible point. □

$\Rightarrow$ Clustering applications:
Information Retrieval
Biology
Business

Summarization
Nearest neighbors
Compression



Introduction Clustering

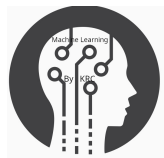
$\Rightarrow$ Consider a group of 12 sales records each indicating sales price, and have been sorted in ascending order as: 5, 8, 11, 13, 15, 35, 45, 55, 72, 92, 201, 215. It is required to partition these into three clusters. The partitions finally formed are shown in Table:

Cluster-1	Cluster-2	Cluster-2
5,8,11, 13,15	35,45,55, 72,92	201,215

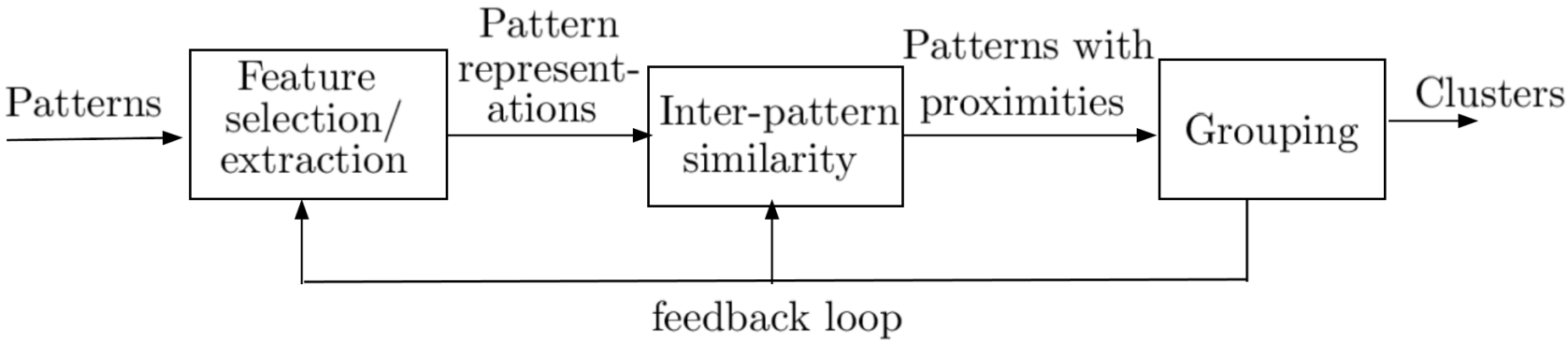
Also, applications are: image segmentation, pattern classification, and data mining.

A general pattern clustering process has following steps:

1. pattern representation, which may also includes feature selection and extraction,
2. defining proximity measures patterns specific to data domains,
3. grouping of patterns (clustering),
4. optionally, abstraction of data, and
5. optionally, assessment of output.



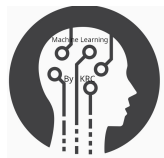
Clustering stages



⇒ The Fig. above is a typical case of clustering. Feedback path indicates that the grouping process output could affect feature extraction and similarity computations in the next iteration.

⇒ The *pattern representation* depends on: available patterns, classes and their number, feature types and their scale for clustering algorithm. Feature selection helps in identifying the most effective subset of the original features to use in clustering.

⇒ Through *feature extraction*, one or more transformations of input features is carried out.



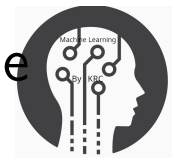
Clustering stages...

$\Rightarrow$ A *Pattern* set is denoted by $\mathscr{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$. The ith pattern in $\mathscr{X}$ is: $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})$. Often, a pattern set to be clustered is viewed as an $n \times d$ pattern matrix.

$\Rightarrow$ *Pattern proximity* or closeness of one pattern to other, is measured by a distance function defined on a pair of patterns. A distance measuring function is *Euclidean distance*: used to reflect dissimilarity between two patterns. And, if the Euclidean distance is zero...

$\Rightarrow$ The *grouping* step can be carried out in many ways. The output of clustering can be *crisp* or *fuzzy*. When clustering at output is crisp (*hard*), the data is partitioned into groups, where as when it is fuzzy partition, each pattern has a variable degree of membership of $[0, 1]$, in each of the output clusters.

$\Rightarrow$ *Fuzzy clustering*. It is procedure that assign to each input pattern $\mathbf{x}_i$ a fractional degree of membership $f_{i,j}$ in each output cluster j, for all the k clusters.



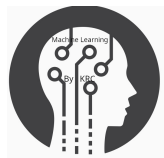
Clustering stages...

$\Rightarrow$ All clustering algorithms produce clusters when presented with data, irrespective of whether the data really contain clusters or not.

$\Rightarrow$ It is not necessary that every set of data contains some clusters. For example, the continuous sequence $1, 2, ..., 100$, in no way represents a cluster. Only what we can have is all these numbers as clusters, each of size one.

$\Rightarrow$ To evaluate the cluster quality, we should know what characterizes a "good" clustering result. If data contain no clusters, some clustering algorithms may obtain "better" clusters than others.

$\Rightarrow$ Criteria for comparing clustering algorithms are based on: 1. how the clusters are formed, 2. their data structure, and 3. how sensitive is the clustering technique to changes, which does not affect the data structure.



Data Clustering and Cluster Analysis

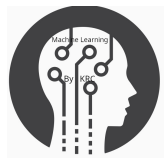
⇒ Clustering is used an initial step in many data mining processes. Some area that use clustering are: *predictive modeling, database segmentation* and *visualization* of large databases.

⇒ Data mining is carried on relational databases: transactions with well-defined structure: columns as features. Also used on large *unstructured* databases: WWW, here content is NL text: HTML /XML.

⇒ Old clustering methods: Complete data should fit into RAM, focus was on clustering quality and not scalability. Old clustering algorithms are not scalable.

⇒ Clustering is used in data-mining for segmenting databases into homogeneous groups, for data compression.

⇒ Clustering helps to identify characteristics of sub-populations that are targeted for specific purposes (e.g., marketing certain items aimed at specific section of population).



⇒ The table 1 (page no. 2) is a sample database as clusters, here clusters of customers' database based on three purchase behavior: quantity, unit price, and their combination.

⇒ We organize a collection of patterns into groups based on their similarity. Intuitively, the patterns in same cluster are more similar to each other than those in different clusters. See Fig. 1 (a) and (b). Measurement of the patterns (x, y) coordinate values.

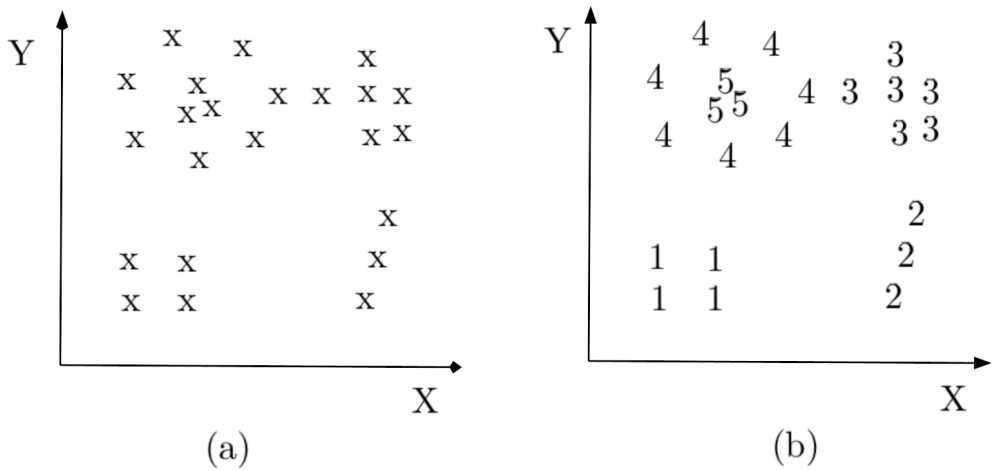


Figure 1: Data Clustering: (a) Input patterns, (b) Clusters formed

Data Clustering and Cluster Analysis

⇒ One technique for clustering is called *supervised learning*, while other is called *unsupervised learning*.

⇒ In supervised technique, a collection of labels, i.e., pre-classified patterns, are already provided (training patterns), and the task is to label newly encountered unlabeled patterns.

⇒ In the case of unsupervised learning, the task is to group a given collection of unlabeled patterns into some meaningful clusters. In fact, some kind of labels are there, associated with the clusters this time also, but this category of labels are data driven – obtained solely from the data, and not predefined.

Chowdhary, K.R. (2020). Data Mining. In: Fundamentals of Artificial Intelligence. Springer, New Delhi. https://doi.org/10.1007/978-81-322-3972-7_17 pp. 519-534.