Clustering in Expert System
Ms. Sangeeta Ahuja

Clustering
Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”.

A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. We can show this with a simple graphical example:

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are “close” according to a given distance (in this case geometrical distance). This is called distance-based clustering. Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

The Goals of Clustering
So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data. But how to decide what constitutes a good clustering? It can be shown that there is no absolute “best” criterion which would be independent of the final aim of the clustering. Consequently, it is the user which must supply this criterion, in such a way that the result of the clustering will suit their needs.
For instance, we could be interested in finding representatives for homogeneous groups (data reduction), in finding “natural clusters” and describe their unknown properties (“natural” data types), in finding useful and suitable groupings (“useful” data classes) or in finding unusual data objects (outlier detection).

Requirements

The main requirements that a clustering algorithm should satisfy are:

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- interpretability and usability.

Problems

There are a number of problems with clustering. Among them:

- current clustering techniques do not address all the requirements adequately (and concurrently);
- dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- the effectiveness of the method depends on the definition of “distance” (for distance-based clustering);
- if an obvious distance measure doesn’t exist we must “define” it, which is not always easy, especially in multi-dimensional spaces;
- the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

Clustering Algorithms

Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. A simple example of that is shown in the figure below, where the separation of points is achieved by a straight line on a bi-dimensional plane.
On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value.

Instead, a hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, the last kind of clustering use a completely probabilistic approach. There are four most commonly used clustering algorithms:

- K-means
- Fuzzy C-means
- Hierarchical clustering
- Mixture of Gaussians

Each of these algorithms belongs to one of the clustering types listed above. So that, K-means is an exclusive clustering algorithm, Fuzzy C-means is an overlapping clustering algorithm, Hierarchical clustering is obvious and lastly Mixture of Gaussian is a probabilistic clustering algorithm. We will discuss about each clustering method in the following paragraphs.

**Distance Measure**

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading. Figure shown below illustrates this with an example of the width and height measurements of an object. Despite both measurements being taken in the same physical units, an informed decision has to be made as to the relative scaling. As the figure shows, different scalings can lead to different clusterings.
Notice however that this is not only a graphic issue: the problem arises from the mathematical formula used to combine the distances between the single components of the data feature vectors into a unique distance measure that can be used for clustering purposes: different formulas lead to different clusterings. Again, domain knowledge must be used to guide the formulation of a suitable distance measure for each particular application.

**Minkowski Metric**

For higher dimensional data, a popular measure is the Minkowski metric,

$$d_p(x_i, x_j) = \left( \sum_{k=1}^{d} |x_{i,k} - x_{j,k}|^p \right)^{1/p}$$

where $d$ is the dimensionality of the data. The *Euclidean* distance is a special case where $p=2$, while *Manhattan* metric has $p=1$. However, there are no general theoretical guidelines for selecting a measure for any given application.

It is often the case that the components of the data feature vectors are not immediately comparable. It can be that the components are not continuous variables, like length, but nominal categories, such as the days of the week. In these cases again, domain knowledge must be used to formulate an appropriate measure.

**Applications**

**Biology**

In biology clustering has many applications
In the fields of plant and animal ecology, clustering is used to describe and to make spatial and temporal comparisons of communities (assemblages) of organisms in heterogeneous environments; it is also used in plant systematics to generate artificial phylogenies or clusters of organisms (individuals) at the species, genus or higher level that share a number of attributes.

In computational biology and bioinformatics:

In transcriptomics, clustering is used to build groups of genes with related expression patterns (also known as coexpressed genes). Often such groups contain functionally related proteins, such as enzymes for a specific pathway, or genes that are co-regulated. High throughput experiments using expressed sequence tags (ESTs) or DNA microarrays can be a powerful tool for genome annotation, a general aspect of genomics.

In sequence analysis, clustering is used to group homologous sequences into gene families. This is a very important concept in bioinformatics, and evolutionary biology in general. See evolution by gene duplication.

In high-throughput genotyping platforms clustering algorithms are used to automatically assign genotypes.

In QSAR and molecular modeling studies as also chemoinformatics

**Medicine**

In medical imaging, such as PET scans, cluster analysis can be used to differentiate between different types of tissue and blood in a three dimensional image. In this application, actual position does not matter, but the voxel intensity is considered as a vector, with a dimension for each image that was taken over time. This technique allows, for example, accurate measurement of the rate a radioactive tracer is delivered to the area of interest, without a separate sampling of arterial blood, an intrusive technique that is most common today.

**Market Research**

Cluster analysis is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers.

- Segmenting the market and determining target markets
- Product positioning
- New product development
- Selecting test markets (see : experimental techniques)
Educational Research

In educational research analysis, data for clustering can be students, parents, sex or test score. Clustering is an important method for understanding and utility of cluster in educational research. Cluster analysis in educational research can be used to find out data exploration, cluster confirmation and hypothesis testing. **Data exploration** is used when there is little information about which schools or students will be grouped together. It aims at discovering any meaningful clusters of units based on measures on a set of response variables. **Cluster confirmation** is used for confirming the previously reported cluster results. **Hypothesis testing** is used for arranging cluster structure.

Example of Cluster Analysis in Educational Research

In 2002, Hattie used cluster analysis in the project 'School Like Mine' to compare students’ achievement in literacy and numeracy by the type of school they attended. 2707 majority and minority students in New Zealand are classified into different clusters according to school size, student ethnicity, region, size of civil jurisdiction and socioeconomic status for comparison. The cluster in this research is calculated across five dimensions, decile, region, size, minority and rurality. All schools are placed into one of twenty clusters that are used in the as TTle software as a basis of student achievement comparison. The result shows that using the power of socioeconomic status to describe schools is analysed and found inadequate. Euclidean distance was used as the clustering method in this research. Schools that are alike were clustered together. In addition, dendrogram. By clustering schools, Hattie suggested that school types had no significant relation with performance of schools.

Advantages of Cluster Analysis

Frisvad of BioCentrum-DTU said that cluster analysis is a good way for quick review of data, especially if the objects are classified into many groups. In the above example, ‘Schools Like Mine, 23 clusters of school with different properties were clearly clustered. Cluster Analysis is easy for user to assign or nominate themselves in to a cluster they would most like to compare with in exist school cluster database because each cluster is clearly named with understandable terms.
Cluster Analysis provides a simple profile of individuals. Given a number of analysis units, for example school size, student ethnicity, region, size of civil jurisdiction and social economic status in this example, each of which is described by a set of characteristics and attributes. Cluster Analysis also suggests how groups of units are determined such that units within groups are similar in some respect and unlike those from other groups.

Disadvantages of Cluster Analysis

Object can be assigned in one cluster only. For example in 'Schools Like Mine', schools are automatically assigned into the first twenty-two clusters. However, if schools want to compare themselves with integrated schools, they will have to manually assign themselves into cluster twenty-three. Data-driven clustering may not represent the reality because once a school is
assigned to a cluster, it cannot be assigned to another one. Some schools may have more than one significant property or fall on the edge of two clusters. Clustering may have detrimental effects to teachers who work in low-decile schools, students who are educated in them, and parents who support them, by telling them the schools are classified as ineffective, when in fact many are doing well in some unique aspects that are not sufficiently illustrated by the clusters formed.

In k-means clustering methods, it is often requires several analysis before the number of clusters can be determined. It can be very sensitive to the choice of initial cluster centres.

Other applications

- Social network analysis
  In the study of social networks, clustering may be used to recognize communities within large groups of people.
- Software evolution
  Clustering is useful in software evolution as it helps to reduce legacy properties in code by reforming functionality that has become dispersed. It is a form of restructuring and hence is a way of directly preventative maintenance.
- Image segmentation
  Clustering can be used to divide a digital image into distinct regions for border detection or object recognition.
- Data mining
  Many data mining applications involve partitioning data items into related subsets; the marketing applications discussed above represent some examples. Another common application is the division of documents, such as World Wide Web pages, into genres.
- Search result grouping
  In the process of intelligent grouping of the files and websites, clustering may be used to create a more relevant set of search results compared to normal search engines like Google. There are currently a number of web based clustering tools such as Clusty.
- Slippy map optimization
  Flickr's map of photos and other map sites use clustering to reduce the number of markers on a map. This makes it both faster and reduces the amount of visual clutter.
- IMRT segmentation
  Clustering can be used to divide a fluence map into distinct regions for conversion into deliverable fields in MLC-based Radiation Therapy.
- Grouping of Shopping Items
  Clustering can be used to group all the shopping items available on the web into a set of unique products. For example, all the items on eBay can be grouped into unique products. (eBay doesn't have the concept of a SKU)
- Recommender systems
  Recommender systems are designed to recommend new items based on a user's tastes. They sometimes use clustering algorithms to predict a user's preferences based on the preferences of other users in the user's cluster.
- Mathematical chemistry
To find structural similarity, etc., for example, 3000 chemical compounds were clustered in the space of 90 topological indices.

- **Petroleum Geology**
  Cluster Analysis is used to reconstruct missing bottom hole core data or missing log curves in order to evaluate reservoir properties.

- **Physical Geography**
  The clustering of chemical properties in different sample locations.

- **Crime Analysis**
  Cluster analysis can be used to identify areas where there are greater incidences of particular types of crime. By identifying these distinct areas or "hot spots" where a similar crime has happened over a period of time, it is possible to manage law enforcement resources more effectively.

- **Evolutionary algorithms**
  Clustering may be used to identify different niches within the population of an evolutionary algorithm so that reproductive opportunity can be distributed more evenly amongst the evolving species or subspecies.

So, it may be concluded that many data mining techniques have been used in agriculture and in expert system. Clustering has been widely used data mining techniques over agriculture data sets. It is a contemporary technique to find the solution over the traditional and conventional method. It is an infant area with respect to data mining application over agriculture.