CSCE 970: Pattern Recognition	Spring Semester, 1999

Topic 7: Clustering:Basic Concepts

Lecturer: Stephen Scott *Scribe: Gautam Jindal*

1 Introduction

In this lecture we will discuss the following:
2) Review of supervised and unsupervised paattern recognition from lecture 1
3) Goal of clustering
4) Steps involved while clustering
5) Applications of clustering
6) Types of features used while clustering
7) Clustering types
8) Definations of proximity measure

2 Review of supervised and unsupervised pattern recognition

In all the previous chapters we were dealing with Supervised Classification i.e we assumed that the set of training data were available and the classifier was designed by exploiting this a priori information. However sometimes the training data, of known class labels are not available, or even more what if the labels itself are unavailable? In these types of problem we unravel the simmilarities between the given feature vectors and cluster them together to form a group. This is known as unsupervised pattern recognition or clustering. An example is multispectral remote sensing, this is the same example covered in lecture 1. This example is demonstrated in the figure1 below.

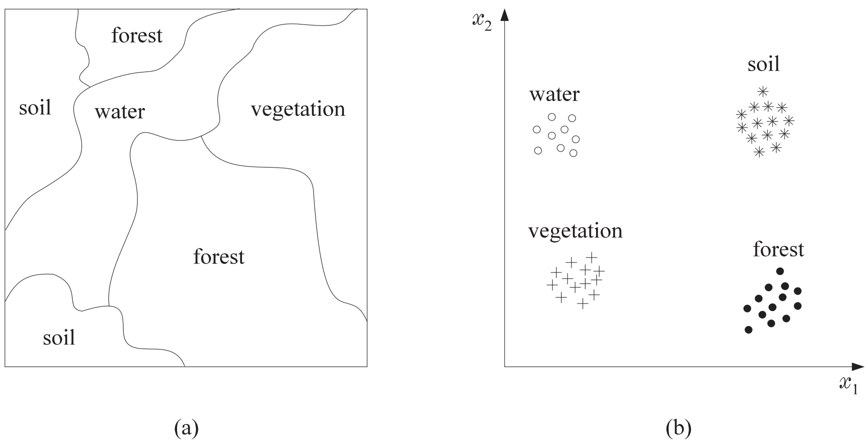


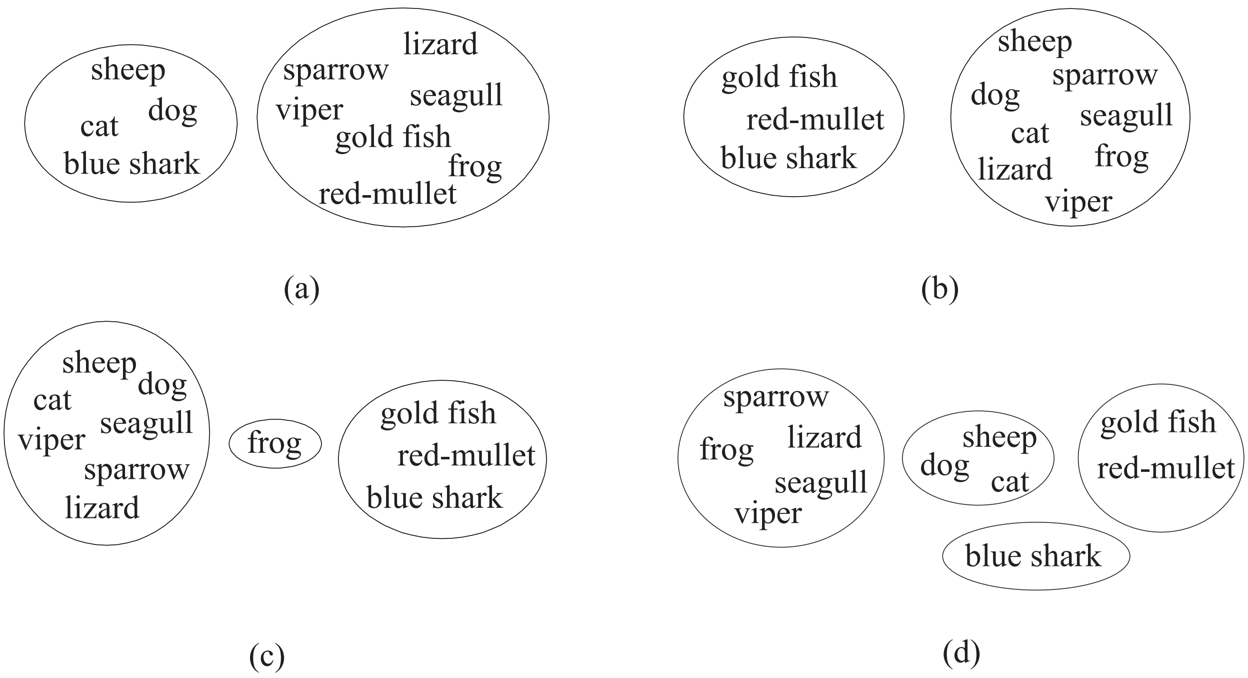
Figure 1: a)An illustration of various types of ground cover and b) clustering of the respective features for multispectral imaging using two bands.

Here the feature vectors are measurements of electromagnetic energy reflected from remote parts of earth, parts where it is not affordable to visit to determine the label. A clustering algorithm can then be used to reveal the groups involved using these feature vectors. Points that correspond to the same ground cover type, such as water, are expected to group together. Once this is done i.e assuming that the clustering is good, then an expert can label each group with the available references either using maps or visiting the site.

3 Goal of clustering

The goal of the clustering (a.k.a unsupervised pattern recognition) algorithm is to place patterns (feature vectors) into "sensible" clusters such that they can be used to reveal similarities and dissimilarities between them, in other words we should be able to derive 'useful' conclusions.

The definations of "sensible" and "useful" will depend upon the application being solved and also involve human intervention. As an example consider the following animals: sheep, dog, cat (mammals), sparrow, seagull (birds), viper, lizard (reptiles), goldfish, red mullet, blue shark (fish) and frog (amphibian). Depending on the application or what we are trying to achieve the clustering algorithm groups the animals above differently under different clustering criterion.This can be observed in the figure2 below.



.

Figure 2: (a) How they bear their young (b) Existence of lungs (c) Environment in which they live (d) Both (a) and (b) (e) [not shown] Vertebrates (all fall in the same cluster

4 Steps involved while clustering

The basic steps an expert should follow in order to develop a clustering task are the following:

4.1 Feature selection

Features with minimum information redundancy is a major goal. But at the sametime features which discriminate the task of interests should always be present. Feature selection requires preprocessing of data. Feature selection was discussed in topic 5, some feature selection methods are Feature selection based on statistical hypothesis, ROC Curve,etc.

All these procedures are also applicable for this step in clustering.

4.2 Proximity measure

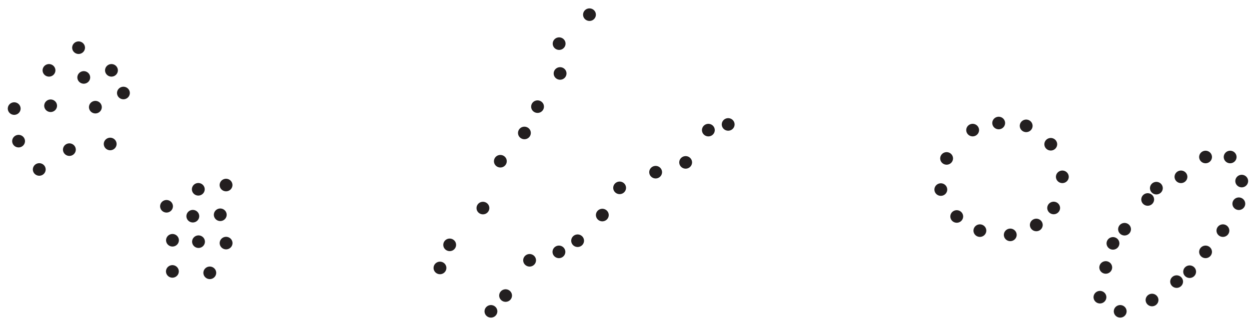
This is a measure that quantifies how "similar" or "dissimilar" two feature vectors are. It is natural to ensure that all features be given equal weightage while computing the proximity measure and no features dominates over the others. This is generally taken care of while preprocessing. Proximity measure can be measured between:

1) feature vectors - compare two feature vectors to decide whether to group or not.

2) feature vectors and a set(cluster) - compare a feature vector to a cluster to decide whether the feature vector belongs to this cluster or not.

3) sets(cluster) - compare two clusters to get a larger cluster.

Quantifying the terms "similar" and "dissimilar" depends very much on the types of clusters involved. Different measures yield different types of clusters, for example similarity measure yields compact clusters, a different measure yields elongated cluster and different ones Spherical and Ellipsoidal. The type of cluster formed also depends upon the clustering criterion. The figure3 below shows few different types of clusters.



(a) (b) (c)

.

Figure 3: (a) compact clusters (b) Elongated clusters (c) Spherical and ellipsodial clusters

4.3 Clustering criterion

This depends on the interpretation that an expert gives to the term "sensible" which inturn depends upon the application and the type of clustering expected from the data. For example a compact cluster mighth be sensible according to one criterion, whereas an elongated one might be sensible according to another. The clustering criterion is expressed using a cost function or some other well defined rules.

4.4 Clustering tendency

This is an important step in which various tests are made to indicate whether the data we have has the tendency to group together or not. For example, we might have a data set which may be completely random in nature and clustering them would be meaningless.

4.5 Clustering Algorithm

The next step after choosing the proximity measure and the clustering criterion is to choose a good clustering algorithm that unravels the clustering of the data set. Clustering algorithms can be Sequential, Hierarchical, Based on cost function optimization, Probabilistic, Fuzzy etc. These will be discussed later in more detail.

4.6 Cluster Validataion

Once the results from the clustering algorithm have been obtained we need to verify how well we have done. That is, we need to find out whether the choice of the clustering algorith's parameters and the clustering type matches the underlying data's clustering structure, whether the mathching classifier has overfitted, etc. All this is acheived using appropiate tests.

4.7 Interpretation

Finally an expert in the aplication field will interpret the results obtained using its own experimental analysis to draw conclusions.

It can be observed that by choosing different features, clustering criterion, proxity measure and clustering algorithm, we can get totally different clustering results. In other words each step is subjective and depends on how the experts interpret them. This can be easily observed in figure 4 [next page].

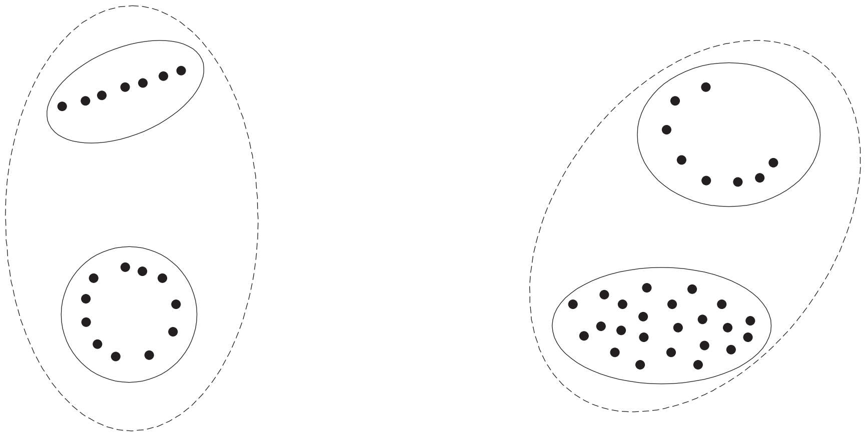


Figure 4: The dashed-line interpretation of the data results in two clusters and the solid-line interpretation results in four clusters

How many "sensilble" ways of clustering can be obtained from the points in figure 4. There are many interpretations, for example if we take the solid line into consideration we get 4 clusters and if we take the dashed line we get 2 clusters. Which one is right? This again depends upon the application we are dealing with and in the end an expert in this application will decide which one is right.

5 Clustering Applications

Clustering have many applications, here we will discuss the four basic applications :

5.1 Data reduction

Many times the data available (N) is very large and requires some compression to reduce the processing demands. By clustering we can group large amount of data into m clusters such that $m \ll N$, doing so we can then choose a single entity to represent this cluster. For example in compact clusters we can choose an entity mean and variance of the distribution to represent the whole cluster. Similarly for elongated ones the end points may represent the whole cluster. Thus data compression is achieved.

5.2 Suggest hypotheses

By clustering data sets into clusters we can then come up with some hypotheses about the nature of the data. In the animal example discussed before we can state an hypotheses that fishes live in water, then we can go about to verify this hypotheses using other data sets.

5.3 Test hypotheses

After stating a hypotheses we need to verify whether it is correct or not. Consider the following hypothesis "Big companies invest abroad". One way to verify this is to apply clustering to large and representative set of companies. A company is represented by correlated features size, activities abroad and its ability to complete projects. If after clustering , a group is formed that is large and has investments abroad will support this hypotheses irrespective to whether they complete the project or not. This shows that hypotheses depends upon certain features that are corelated while others are independent.

5.4 Prediction based on groups

This is helpful in naming or characterizing the cluster after clustering has taken place. A resulting group can be given a characterized name by picking up few feature vectors from this group , after characterization any feature vector falling within this group will be named by this characterized name.

6 Types of features

A feature can be of different types based on the fact whether it takes its value from a continuous range or from a finite discrete set. If the discrete set has only 2 elements then the feature is called dichotomous or binary.

A different characterization is based on the relative signifiacance between each feature of the feature vector. We have four categories, they are as follows:

6.1 Nominal

It is not necessary that data given in real world problems will always be real values , non-quantified data can also exist. Nominal falls into this category where the features are given a name only. For example Gender { male, female } , it is clear that no quantitative comparisions can be made between these features.

6.2 Ordinal

Ordinal takes one step further, in this case the features are not only named but it can be valued so as to get them "meaningfully ordered". The difference between two successive values is of no meaningful "quantitative" importance as the value is given only to order the features. For example we can give the values {4,3,2,1} to the names corresponding to the rating "excellent", "very good", "good" and "not good" ,thus forming a meaningful order. We could also order them by giving the values {0.4,0.3,0.2,0.1} thereby showing that the quantitative difference has no meaning.

6.3 Interval-scaled

Interval-scaled inherits from ordinal, i.e. not only can the features be meaningfully ordered, the difference between two feature values is also meaningful but their ratios are meaninigless. For example consider temperature. If the temperatures in Lincoln and Chicago are 80 degrees and 40 degrees respectively, then it is meaningful to say that the temperature in Lincoln is 40 degrees higher than that in Chicago. However it is meainingless to say that Licoln is twice as hot as Chicago.

6.4 Ratio-scaled

Ratio-scaled inherits from Interval-scaled. In this case both, the difference between two feature values and the ratio between them are meaningful. For example weight, it is meaningful to say that a person who weighs 100kg weighs twice as much as a person who weighs 50kg, and also that this persons weight is 50kg's more.

From the explanation given above about the types of features, it can be easily observed that each type of feature posssess all the properties of the type before it. In other words Ordinal inherits from Nominal, Interval-scaled from Ordinal , and Ratio-scaled from Interval-scaled.

7 Clustering types

Many definations are proposed for clustering over the years [John 67,Ever 81], but most of these definations are based on terms such as "similar", "alike", etc. It was pointed out in [Ever 81] that most of these defiantions are vague and circular in nature. The fact being that there is no universally acceptable defination for clustering. For our usage we shall define clustering as follows, given a data set X = { $x_1, x_2, x_3, ...x_n$ } we are interested in placing them into m clusters $C_1, C_2, C_3....C_m$.

Two types of Clustering are as follows:

7.1 Hard(Crisp)

In hard clustering, the data set X = { $x_1, x_2, x_3, ...x_n$ } is placed into m clusters in such a way that set X forms a partition of m sets $C_1, C_2, C_3....C_m$. The following conditions are to be met to achieve this:

1) $C_i \neq 0$

This condition makes sure that out of the m clusters formed no cluster will be empty, i.e. without any feature vector.

2)

$$\bigcup_{i=1}^{m} = X$$

This condition makes sure that we do not exclude any feature vector from the data set X. In other words each and every feature vector in data set X will be assigned a cluster.

3)$C_i \cup C_j = 0$, i $\neq$ j, i, j $\in$ { 1,....m }

This condition makes sure that no two clusters share the same feature vector, i.e. each feauture vector will belong to only one cluster and only one.

In addition to these conditions we are required to hold another property, the feature vectors belonging to cluser i should be "more similar" to each other and "less similar" to all the other clusters. To quantify the terms "similar" and "less similar" we need to decide upon the proximity measure. Different proximity measures lead to different types or shapes of clusters as discussed above.

7.2 Fuzzy

Fuzzy clustering was introduced by Zadeh[Zade 65]. In fuzzy clustering the data set X is placed into m clusters based on the membership functions μ_j where

μ_j : X $\rightarrow$ [0,1], j=1,....,m

In other words the function takes in a feature vector and maps it to a real value between 0 and 1. What we are trying to do here is we take a feature vector and try to put part of it in $C_1, C_2, .., C_j$ thereby definig the degree of membership the feature vector holds in each cluster. If the degree of membership of a feature vector for a particular cluster C_i is 1 then we say that the feature vector belongs 100and if the degree of membership of a feature vector for a particular cluster C_i is 0 then we say that the feature vector does not belong to cluster C_i.

The other conditions to be met are,

1)

$$\sum_{j=1}^{m} \mu_j x_i = 1, i \in \{1, 2, ...N\}$$

2)

$$0 < \sum_{i=1}^{N} \mu_j x_i < N, j \in \{1, 2, ...m\}$$

These two conditions guarantees that there are no empty clusters and not all the features lie in one single cluster.

If two feature vectors x_k, x_n have a high membership value for a particular cluster then they are considered similar to each other [Wind 82].

If the fuzzy membership functions μ_j takes values from set $\{0, 1\}$, ie either 0 or 1, then each vector will exclusively belong to one cluster only and the membership functions are known as characteristic functions [Klir 95].

8 Definations of Proxity Measure

8.1 Dissimilarity measure

Dissimilarity measure (DM) d on X is a function defined below

d : X $\times$ X $\rightarrow \Re$

Here X is defined as a feature vector and $\Re$ a set of real numbers. In other words d is a function that takes in two feature vectors and returns a quantitative real value $\Re$.

$\Re$ is defined by a finite set of values which ranges such that

$\exists\ d_0 \in \Re$: $-\infty < d_0 \leq$ d(x,y) $< +\infty$, $\forall$ x, y$\in$ X

d(x,x) $= d_0\ \forall$ x $\in$ X

d(x,y) = d(y,x) $\forall$ x, y$\in$ X

d_0 is the lowerbound which symbolizes the smallest dissimilarity value generally 0.

If in addition to the above rules, the following two rules hold good then d is called a metric DM

$d(x,y) = d_0 \iff x = y$
Whatever be the smallest dissimilarity value the only way it can be achieved is by d(x,x), the dissimilarity measure of the feature vector by itself.

$d(x,z) \leq d(x,y) + d(y,z), \forall\ x, y, z \in X$
This equation is known as the triangular inequality.

A good example of a dissimilarity measure is $d(\cdot,\cdot)$, which is the Euclidean distance between two points. Here we can see that the lowerbound $d_0 = 0$. $d(\cdot,\cdot)$ follows all the above rules including the traingular inequality rule thus $d(\cdot,\cdot)$ is a metric DM.

8.2 Similarity measure

Similarity measure is basically a dual of the dissimilarity measure. There are ways to convert these measures into one another. In similarity measure we are concerned of how similar the vectors are , if the similarity measure is high then we can group the two concerned vectors into a single cluster else group them into different clusters.

Similarity measure (SM) s on X is a function defined below
$s : X \times X \rightarrow \Re$

Here X is defined as a feature vector and $\Re$ a set of real numbers. In other words d is a function that takes in two feature vectors and returns a quantitative real value $\Re$.

$\Re$ is defined by a finite set of values which ranges such that

$\exists\ s_0 \in \Re$: $-\infty < s(x,y) \leq s_0 < +\infty, \forall x,y \in X$

$s(x,x) = s_0\ \forall\ x \in X$

$s(x,y) = s(y,x)\ \forall\ x, y \in X$

s_0 is the upperbound.

If in addition to the above rules, the following two rules hold good then s is called a metric SM

$s(x,y) = s_0 \iff x = y$
Whatever be the upperbound value the only way it can be achieved is by s(x,x), the similarity measure of the feature vector by itself.

$s(x,y)s(y,z) \leq [s(x,y) + s(y,z)]\ s(x,z), \forall\ x, y, z \in X$

8.3 Proximity measures between sets

In the previous two sections we were dealing with proximity measure between two feature vectors, now we shall extend our measure between two sets of feature vectors. This type of measure is usually used in hierarchical clustering algorithm, which we shall discuss later.

In this case let U = $\{ D_1, D_2, ..., D_k \}$, $D_i \subset$ X. In other words U is a set which contains the subsets of of X. the proximity measure α on U is a function defined as follows

α : U $\times$ U $\rightarrow \Re$

All the equations used in dissimilarity measure and similarity measure can be used now with D_i,D_j in place of x and y and U in the place of X.

Generally when we are measuring the proximity measure between two sets D_i and D_j we do so by measuring the proximity measure between the elements belonging to the sets D_i and D_j respectively.

Take for example X = $\{ x_1, x_2, x_3, x_4, x_5, x_6 \}$

U = $\{ \{ x_1, x_2 \}, \{ x_1, x_4 \}, \{ x_3, x_4, x_5 \}, \{ x_1, x_2, x_3, x_4, x_5 \} \}$

Let us also define the following dissimilarity measure

d^{ss}_{min} $(D_i, D_j) = min_{x \in D_i, y \in D_j}$ D_2 (x,y)

where d_2 is the Euclidean distance between elements D_i and $D_j \in U$.

Also the minimum possible value of d^{ss}_{min} is $d^{ss}_{min,0} = 0$ which is equal to d^{ss}_{min} (D_i, D_i), this is the same as the other measures.

Also if d^{ss}_{min} (D_i, D_j)=d^{ss}_{min} (D_j, D_i) then d^{ss}_{min} $(\cdot,\cdot)$ is a DM.

However as d^{ss}_{min} $(\{ x_1, x_2 \}, \{ x_1, x_4 \}) = d^{ss}_{min,0}$ using the equations above and $\{ x_1, x_2 \} \neq \{x_1, x_4 \}$, this measure is not a metric DM.

References

Following are the references cited in this topic:

[Ever 81] Everitt B. Cluster Analysis, 2nd ed., Halsted Press, 1981.

[John 67] Johnson S.C. "Heirarchical clustering schemes", Psychometrika, Vol. 32, pp.241-254, 1967

[Klir 95] Klir G., Yuan B., Fuzzy sets and fuzzy logic, Prentice Hall, 1995

[Wind82] Windham M.P. "Cluster validity for fuzzy c-means clustering algorithm", IEEE transactions on Pattern Analysis and Machine Intelligence, Vol. 4(4),pp.357-363, 1982

[Zade 65] Zadeh L.A. "Fuzzy sets", Information and Control,Vol. 8,pp.228-353, 1965.