CSCE 970: Pattern Recognition	Spring Semester, 1999

Topic 7: Clustering:Basic Concepts

Lecturer: Stephen Scott *Scribe: Luyin Cao*

1 Introduction

In previous lectures we gave the definition of dissimilarity measure and similarity measure. Generally we choose the Euclidean distance as a measure of dissimilary. In similarity measure we focus on how similar the features are . We also extended the definitions to measure "proximity" between subsets of $\mathbf{x}$. In this leature we will talk about the proximity measure between two points, the proximity measure between a point and a set, the proximity measure between two sets. And we will also give an overlook of clusting algorithms at the end of this lecture.

2 Proximity Measure between Two Points

In order to develop a clusting task, processing of features may be necessary prior to their utilization in subsequent stages. Proximity measure is a measure that quantifies how "similar" or "dissimilar" two feature vectors are. Based on the definitions of proximity meaures, we give the proximity measure between two points at first.

2.1 Proximity Measure between Points with Real-valued Vectors

2.1.1 Example of Dissimilarity Measures

In practice, the general purpose metric DMs between real-valued vectors are:

- The weighted l_p metric DMs

  $$d_p(\mathbf{x}, \mathbf{y}) = (\sum_{i=1}^{l} \omega_i |x_i - y_i|^p)^{l/p}$$

  where x_i, y_i are the ith coordinates of $\mathbf{x}$ and $\mathbf{y}$, i=1,...,l, and $\omega_i \geq 0$ is the ith weighted coefficient, which are used mainly on the real- valued vectors.

- Special cases of metric DMs:

  If $\omega_i = 1$, and $p = 2$, the unweighted l_p metric DM is the Eulidean distance.

  If $p = 1$, the weighted Manhattan distance is:

  $$d_1(x, y) = \sum_{i=1}^{l} \omega_i |x_i - y_i|$$

  If $p = 2$, the generalization of the weighted l_2 norm can also be:

  $$d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T B (\mathbf{x} - \mathbf{y})}$$

  where B is a symmetric, positive definition matrix. Mahalanobis distance is a special case of it.

2.1.2 Example of Similarity Measures

The most important common similarity measures for real-valued used in practice are:

- The inner product:

  It is defined as $Sinner(\mathbf{x}, \mathbf{y}) = \mathbf{x^T y} = \sum_{i=1}^{l} x_i y_i$

  Generally, we normalize the vector $\mathbf{x}$ and $\mathbf{y}$ when we use the inner product. That is, if $\|\mathbf{x}\|_2, \|\mathbf{y}\|_2 \leq a$, $Sinner(\mathbf{x}, \mathbf{y})$ should satisfy the inequility:

$$-a^2 \leq Sinner(\mathbf{x}, \mathbf{y}) \leq a^2$$

  Intuitively speaking, inner product may be considered as a measure of the correlation between x and y.

- Tanimote distance:

$$S_T(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x^T y}}{\|\mathbf{x}\|^{\mathbf{2}} + \|\mathbf{y}\|^{\mathbf{2}} - \mathbf{x^T y}}$$

  $S_T(\mathbf{x}, \mathbf{y})$ is called Tanimoto measure which is also knowen as Tanimoto distance. It is also used for discreted-valued vectors. After some algebraic manipulation, we can obtain:

$$S_T(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{1}}{\mathbf{1} + \frac{\mathbf{(x-y)^T(x-y)}}{\mathbf{x^T y}}}$$

  Hence, we also say that $S_T(\mathbf{x}, \mathbf{y})$ is inversely proportional to the squared Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$ devided by their correlation. that is, the more correlated $\mathbf{x}$ and $\mathbf{y}$, the large the value of S_T.

2.2 Proximity Measure between Two Points Based on Discrete-valued Vectors

If the coordinates of feature vectors come from the finite set $F = 0, \ldots, k-1$, where k is a position interger. We can use DMs and SMs which are defined for real-valued feature vectors.

Dissimilarity Measures:

Here we use Hammimg distance which is defined as the number of places where two victors differ. The Hamming distance d_H is $d_H(\mathbf{x}, \mathbf{y}) = \sum_{\mathbf{i=1}}^{\mathbf{k-1}} \sum_{\mathbf{j=0, j \neq i}}^{\mathbf{k-1}} \mathbf{a_{ij}}$.

So the DMs indicate the number of places where x and y differ.

Similarity Measure:

We applied the Tanimoto measure as the similarity measure for discret-valued. It is defined as

$$\frac{n_{x \cap y}}{n_x + n_y - n_{x \cap y}} = \frac{n_{x \cap y}}{n_x \cup y}$$

where $n_x, n_y, n_{x \cap y}$ are the cardinalities of vectors x,y and $x \cap y$.

So SM measures the radio of the number of places where x and y are the same to the total number of places.

But when using Tanimoto measure, the measure should ignore these places where x_i and y_i are bith 0.

2.3 Proximity measures between two points with the fuzzy Measures

In this part, we consider real-valued vectors $\mathbf{x}, \mathbf{y}$ whose components x_i,y_i satisfy $x_i \in [0,1]$, ($y_i \in [0,1]$). The value of x_i (y_i) are the measure of how much **x (y)** possesses the ith feature.

- If $x_i, y_i \in [0,1]$, we define the similarity between two variables as

  $$(x_i \equiv y_i) = ((\neg x_i \wedge \neg y_i) \vee (x_i \wedge y_i))$$

  That is, if x_i=y_i=0(1),$((\neg x_i \wedge \neg y_i) \vee (x_i \wedge y_i))$=1. If x_i and y_i are different, x_i=0(1) and y_i=1(0), $((\neg x_i \wedge \neg y_i) \vee (x_i \wedge y_i))$=0.

- Generalizing fuzzy values:

  In the fusszy logic context, the logic "AND" is replaced by the operator min, while the logical "OR" is replaced by the operator max. Also, the logical "NOT" on x_i is replaced by 1-x_i.

  So the degree of similarity between two variables may be defined as:

  $$S(x_i, y_i) = max(min(1 - x_i, 1 - y_i), min(x_i, y_i)).$$

- To measure the similarity between vectors.Now we can define a similarity measure between two vectors $\mathbf{x}, \mathbf{y}$ as

  $$S_F^q(x, y) = (\sum_{i=1}^{l} S(x_i, y_i)^q)^{1/q}$$

  And $S_F^q(x, y)$ satisfies the inequility as

  $$l^{1/q} \leq S_F^q \leq 0.5 l^{1/q}$$

$l^{1/q}$ and 0.5 $l^{1/q}$ are the maxmum and minimum values of S_F respectively.

As q $\to \infty$ we get

$$S_F^{\infty} = max_{1 \leq i \leq l} S(x_i, y_i)$$

As q=1,we get

$$S_F(x, y) = \sum_{i=1}^{l} S(x_i, y_i)$$

The value of $S_F^q(x, y)$ depends not only on the relative position of **x** and **y** in the vector space, but also on their closeness to the center of the vector space. At the center of the vector space, the value of $S_F^q(x, y)$ obtain its mininum value.

3 Proximity Measures between a Point and a Set

In many clusting algorithms , we might want to assign a vector **x** to a existing cluster C. The proximity between **x** and C, $\alpha(\mathbf{x}, C)$, can be defined either by using all points of C or by using a representative of C.

3.1 Proximity Measures by Using All Points of C

If all points of C contribute to $\alpha(\mathbf{x}, c)$, we choose :

- *The max proximity function*:

$$\alpha^{ps}_{max}(\mathbf{x}, C) = max_{y \in C}\alpha(\mathbf{x}, \mathbf{y})$$

- *The min proximity function:*

$$\alpha^{ps}_{min}(\mathbf{x}, C) = min_{\mathbf{y} \in C}\alpha(\mathbf{x}, \mathbf{y})$$

- *The average proximity function:*

$$\alpha^{ps}_{avg}(\mathbf{x}, C) = \frac{1}{n_C}\sum_{y \in C}(\alpha(\mathbf{x}, \mathbf{y}))$$

Where n_C is the cardinality of C.

In these definition, $\alpha(\mathbf{x}, \mathbf{y})$ is any proximity measure between two points.

3.2 Proximity Measures by Using a Representative of C

If we choose a representative of C, the proximity measure between $\mathbf{x}$ and C is measured as the proximity between x and the representative of C. Appropriate choice of a representative depends on the type of cluster. Point representatives suit to compact clusters (Figure 1 a) and hyperplane (hyperspherical) is suitable for clusters of linear shape (Figure1 .b) (hyperspherical shape Figure 1 c).

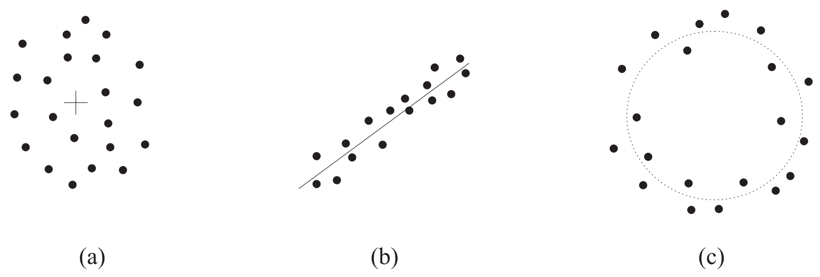

Figure 1: (a)Compact Cluster. (b) Hyperplanar (linear) cluster. (c)Hyperspherical cluster.

3.2.1 Examples of point representatives

Special choices for a point representative of a cluster are :

- Mean vector (mean point)

$$\mathbf{m_p} = \frac{1}{n_C}\sum_{y \in C}\mathbf{y}$$

where n_C is the cardinality of C.

It works well when the point representatives are in a continuous space. but when the points are in a discrete space, it might not exist. For this situation, we can use the mean center $\mathbf{m_c}$ of C.

- Mean center m_c is defined as

$$\sum_{y \in C} d(\mathbf{m_c}, \mathbf{y}) \leq \sum_{\mathbf{y} \in \mathbf{C}} \mathbf{d}(\mathbf{z}, \mathbf{y}), \forall \mathbf{z} \in \mathbf{C}$$

where d is a dissimilarity measure between two points.

If SM is used, the inequality is reversed.

- Median center $\mathbf{m_{med}} \in \mathbf{C}$:

For each point $y \in C$, we find median dissimilarity distance from $\mathbf{y}$ to all other points of C, then take the minimum. So $\mathbf{m_{med}}$ is defined as:

$$med(d(\ bfm_{med}, y) \mid y \in C) \leq med(d(\mathbf{z}, \mathbf{y}) \mid \mathbf{y} \in \mathbf{C}), \forall \mathbf{C}$$

where d is the dissimilary measure between two points.

3.2.2 Examples of Hyperplane & Hypersherical Representatives

- For a linear shaped clusters For linear shaped clusters (or hyperplane, we often choose lines (hyperplanes) as representatives of the clusters. The definition of hyperplane H is :

$$\mathbf{a}^T \mathbf{x} + a_0 = 0$$

where $x = [x_1, \ldots, x_l]$ and $a = [a_1, \ldots, a_l]$ is the weight vector of H.

The distance function of a point $\mathbf{x}$ from H is defined as

$$d(\mathbf{x}, H) = min_{Z \in H} d(\mathbf{x}, \mathbf{z})$$

Using simple geometric arguments ,we can get the Euclidean distance between two points(Figure 2a) is

$$d(\mathbf{x}, H) = \frac{\mid \mathbf{a^T x} + a_0 \mid}{\|\mathbf{a}\|}$$

Where $\|\mathbf{a}\| = \sqrt{\sum_{\mathbf{j=1}}^{\mathbf{l}} \mathbf{a_j^2}}$

- For a hypersphere Q, the definition of it is

$$(\mathbf{x} - \mathbf{c})^\mathbf{T} (\mathbf{x} - \mathbf{c}) = \mathbf{r^2}$$

where r is the radius of the hypersphere and c is the center of it. The distance from a point $\mathbf{x}$ to Q is defined as

$$d(\mathbf{x}, \mathbf{Q}) = \mathbf{min_{Z \in Q} d}(\mathbf{x}, \mathbf{z})$$

Normally, the Euclidean distance between two points is used in this definition (Figure 2b).

- Given a set of points, we can find the representative via regression techniques by minimizing the sum of distance between points and representative.

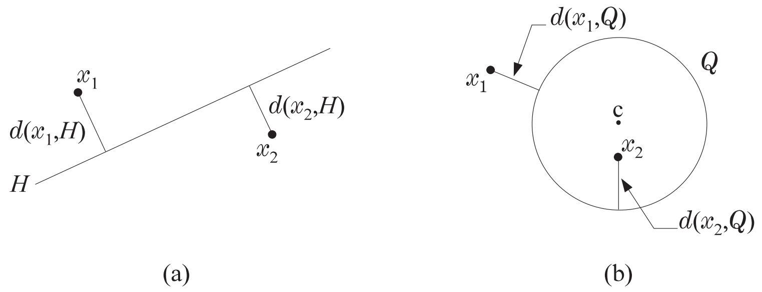

Figure 2: (a)Distance between a point and a hyperplane. (b) Distance between a point and hypersphere.

4 Proximity Measures between Two Sets

Up to now, we have discussed proximity measures between points in l-dimensional spaces and proximity functions between points and sets. Now our goal is to define proximity functions between sets of points. Given two sets of feature vectors D_i and D_j and proximity measure $\alpha(\cdot,\cdot)$ the most common proximity functions are:

The max proximity function:

$$\alpha_{max}^{ss} = max_{x\in D_i, y\in D_j}\alpha(\mathbf{x},\mathbf{y})$$

We can see that if α is a dissimilary measure, α_{max}^{ss} is not a measure. Since it does not satisfy the conditions about the dissimilarity measure. On the other hand, if α is a similarity measure, α_{max}^{ss} is a measure but it is not a metric. So the α_{max}^{ss} is a measure (but not necessarily a metric) iff α is a similarity measure.

Here are examples:

—α is Euclidean distance to measure the dissimilarity between two vectors. We assume l=1, $D_1 = \{(1),(10)\}, D_2 = \{(4),(7)\}, \alpha_{max}^{ss}, (D_1, D_1) = 9 \neq 3 = \alpha_{max}^{ss}(D_2, D_2)$

—If α is a similarity measure, such that

$$\alpha(\mathbf{x},\mathbf{y}) \leq \mathbf{S_0}\forall\mathbf{x},\mathbf{y}$$

and

$$\alpha(\mathbf{x},\mathbf{x}) = \mathbf{S_0}\forall\mathbf{x}$$

So $\alpha_{max}^{ss}(D_i, D_j) \leq S_0, \forall D_i, D_j$,

and

$$\alpha_{max}^{ss}(D,D) = max_{x\in D, y\in D}\{\alpha(\mathbf{x},\mathbf{y})\} = \mathbf{max_{x\in D}}\{\alpha(\mathbf{x},\mathbf{x})\} = \mathbf{S_0}\forall\mathbf{D}$$

We can see α_{max}^{ss} is fully determined by the pair($\mathbf{x}$,$\mathbf{y}$) of the most similar(closest) vectors, with $\mathbf{x} \in D_i$,$\mathbf{y} \in D_j$.

- *The min proximity function*:

$$\alpha_{max}^{ss} = min_{x\in D_i, y\in D_j}\alpha(\mathbf{x},\mathbf{y})$$

In this case when α is a similarity measure, α_{min}^{ss} is not a measure. α_{min}^{ss} is fully determined by the pair($\mathbf{x}$,$\mathbf{y}$) of most dissimilar (distant) vectors, with $\mathbf{x} \in D_i$ and $\mathbf{y} \in D_j$. So α_{min}^{ss} is a measure but it is not a metric iff α is a DM.

- *The average proximity function*

$$\alpha^{ss}_{avg}(D_i, D_j) = \frac{1}{\mid D_i \mid\mid D_j \mid} \sum_{x \in D_i} \sum_{y \in D_j} \alpha(\mathbf{x}, \mathbf{y})$$

  Where $|D_i|$ and $|D_j|$ are the cardinalities of D_i and D_j, respectively. $\alpha^{ss}_{avg}(D_i, D_j)$ is not necessarily a measure even though α is a measure.

- *The Representative (mean):*

  $\alpha^{ss}_{mean}(D_i, D_j) = \alpha(\mathbf{m_{D_i}}, \mathbf{m_{D_j}})$ where m_{D_i} is the represantative of D_i, $i = 1, 2$, $\mathbf{m_{D_i}}$may be the mean point, the mean center, or the median of D_i.

It is obvious that $\alpha^{ss}_{mean}(D_i, D_j)$ is a measure whenever α is a measure.

We should notice that the proximities between two sets are built upon proximities between two points. And the different choices of proximity functions between sets may lead to totally diffient clustering results.

5 Overview of Clustering algorithms

In the previous lectures, we mainly introduce a number of proximity measures. Now our major focus is on the various clustering algorithm.

5.1 Exhausitive search of possible clustings

If we want to find the set of clusters that maximizes SM or minimizes DM, how should we do? One option is to try all possible clusters of size m for various value of m. Let $S(N, m)$ be the number of all possible clusterings of N vectors into m unempty groups, it can be given by the Stirling numbers of hte second kind: $S(N, m) = mS(N - 1, m) + S(N - 1, m - 1)$ It is clear that the brute force approach is infeasible because one has to enumerate all possible clusterings for all possible values of m.

5.2 Categories of Clustering Algorithms

A clustering algorithms is a learning procedure that tries to identify the specific characteristics of the clusters underlying the data set. Clustering algorithms may be devided into the following major categories.

- Sequential algorithms

  These algorithms produces a single clustering. They use straightforward greedy approaches. In most of them, all the feature vectors are presented to the algorithm once or a few times, and the output depends on the order in which the vectors are presented to the algorithm.

- Hierarchical algorithms

  These schemes produce a sequence of clusterings and basically there are two different types:

  —Agglomerative algorithms: These algorithms produce a sequence of clustering of decreasing number of clusters, m at each step. The clustering is produced by repeatedly merging two clusters into one.

—Divisive: These algorithms act in the opposite direction. The clustering is produced at each step by repeatedly dividing cluster into two.

- Algorithms based on cost function optimization

  These algorithms evaluate the goodness of a clustering with a cost function. Typically the number of clusters m is kept fixed.

References

1.Duda R. O., Hart P. *Pattern Classification and Scene Analysis*, Joho Wilry and Sonns,1973.

2.Keinosuke Fukunaga, *Introduction to Statistical Pattern Recognition*, Academic Press, 1990.