Clustering is often called unsupervised learning, because unlike supervised learning, class values denoting an a priori partition or grouping of the data are not given.

4.1 Basic Concepts

Clustering is the process of organizing data instances into groups whose members are similar in some way.

2 reasons for clustering

(1)it's hard to detect clusters visually as dimensions increases

(2)clusters are not clear-cut or well separated

2 types of clustering

(1)partitional clustering

(2)hierarchical ~

Goal of clustering

To discover the intrinsic grouping of the input data through the use of a clustering algorithm and a distance function

Empty clusters problem

choose a data point as the replacement centroid, e.g., a data

point that is furthest from the centroid of a large cluster. If the sum of the

squared error (SSE) is used as the stopping criterion, the cluster with the

largest squared error may be used to find another centroid.

解决方案

如果所有的点在指派步骤都未分配到某个簇，就会得到空簇。如果这种情况发生，则需要某种策略来选择一个替补质心，否则的话，平方误差将会偏大。

（1）选择一个距离当前任何质心最远的点。这将消除当前对总平方误差影响最大的点。

（2）从具有最大SSE的簇中选择一个替补的质心，这将分裂簇并降低聚类的总SSE。如果有多个空簇，则该过程重复多次。

空聚类

[*https://blog.csdn.net/shwan\_ma/article/details/80096408*](https://blog.csdn.net/shwan_ma/article/details/80096408)

4.2.2　Disk Version of the K-means Algorithm

In each for-loop, the algorithm simply scans the data once.

The whole clustering process thus scans the data t times, where t is the

number of iterations before convergence,this algorithm may be used to cluster large data sets which cannot be loaded into the main memory.

Algorithm disk-k-means(k, D) 
Choose k data points as the initial centriods my, j l, 
k; 
2 
3 
4 
5 
6 
7 
8 
9 
10 
Il 
12 
repeat 
initialize sj 0, j l, 
initialize 0, j l, 
for each data point x e D do 
j 4— arg min dist (x, m 
assign x to the cluster j; 
endfor 
until the stopping criterion is met 
0 is a vector with all o's 
// nj is the number of points in cluster j 

Fig. 4.4. A simple disk version of the k-means algorithm

4.2.3　Strengths and Weaknesses

strength

Its time complexity is O(tkn), where n is the number of data points, k is the number of clusters,and t is the number of iterations.Since both k and t are normally much smaller than n. The k-means algorithm is considered a linear algorithm in the number of data points.

weakness

1. The algorithm is only applicable to data sets where the notion of the mean is defined. Thus, it is difficult to apply to categorical data sets.The similarity (or distance) between a data instance and a mode is the number of values that they match (or do not match).

2. The user needs to specify the number of clusters k in advance.

3. The algorithm is sensitive to outliers.

a.Assign each remaining data point to the centroid closest to it.

b.Use the clusters produced from the sample to perform supervised learning (classification).

c.Use the clusters produced from the sample as seeds to perform semisupervised learning.

4. The algorithm is sensitive to initial seeds

a.Each subsequent data point xi is selected such that the

sum of distances from xi to those already selected data points is the largest.

b.sample the data and use the sample to perform hierarchical clustering

c.manually select seeds

5.The k-means algorithm is not suitable for discovering clusters that are not hyper-ellipsoids (or hyper-spheres).

4.3　Representation of Clusters

4.3.1 Common Ways of Representing Clusters

1. Use the centroid of each cluster to represent the cluster.

The centroid representation alone works well if the clusters are of the hyper-spherical shape.

2. Use classification models to represent clusters

a.the situation may not be so ideal.However, there is usually a dominant or large rule which covers most of the data points in the cluster.

b.One can use the set of rules to evaluate the clusters to see whether they conform to some existing domain knowledge or intuition.

3. Use frequent values in each cluster to represent it.

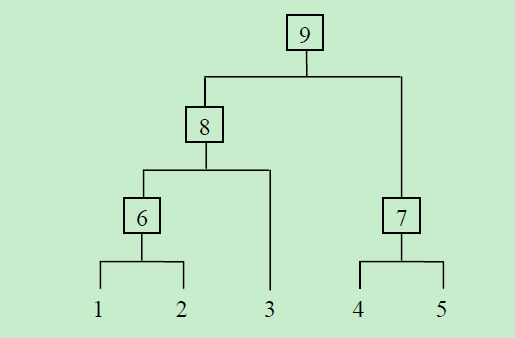
This method is mainly for clustering of categorical data (e.g., in the k-modes clustering).

It is also the key method used in text clustering, where a small set of frequent words in each cluster is selected to represent the cluster.

4.3.2 Clusters of Arbitrary Shapes

4.4 Hierarchical Clustering

Singleton clusters (individual data points) are at the bottom of the tree and one root cluster is at the top, which covers all data points.



There are two main types of hierarchical clustering methods:

1.Agglomerative (bottom up) clustering:

It builds the dendrogram (tree)from the bottom level, and merges the most similar (or nearest) pair of clusters at each level to go one level up. The process continues until all the data points are merged into a single cluster (i.e., the root cluster).

2.Divisive (top down) clustering

It starts with all data points in one cluster,the root. It then splits the root into a set of child clusters. Each child cluster is recursively divided further until only singleton clusters of individual data points remain, i.e., each cluster with only a single point.

Algorithm Agglomerative(D)

1 Make each data point in the data set D a cluster,

2 Compute all pair-wise distances of x1, x2, …, xn ∈ D;

2 repeat

|  |  |
| --- | --- |
| 3 | find two clusters that are nearest to each other; |
| 4 | merge the two clusters form a new cluster c; |
| 5 | compute the distance from c to all other clusters; |

6 until there is only one cluster left

4 
4 
3 
o P3 3,' 
(A). Nested clusters 
2 
P4 
(B) Dendrogram 

Fig. 4.13. The working of an agglomerative hierarchical clustering algorithm

4.4.1 Single-Link Method

合并对象：two clusters with the smallest minimum pair-wise distance.

缺点：it can be sensitive to noise in the data, which may cause the chain effect and produce straggly clusters.

时间：O(n2)

0 0 
,'0 0 
0 0 
0 0 0 
0 

4.4.2 Complete-Link Method

对象：the two clusters with the smallest maximum

pair-wise distance

缺点：it can be sensitive to outliers（still better）

时间：O(n2log n)

4.4.3 Average-Link Method

对象：the two clusters with the smallest average

pair-wise distance

缺点：This is a compromise

时间：O(n2log n)

The following two methods are also commonly used:

Centroid method

Ward's method

4.4.4 Strengths and Weaknesses

Strength

1.able to take any form of distance

or similarity function

2.enables the user to explore clusters at any level of detail

3.It can also find clusters of arbitrary shapes, e.g., using the single-link method.

weakness

1.their computation complexities and space requirements, which are at least quadratic,inefficient and not practical

for large data sets.

2.use sampling to deal with the problems,some scale-up methods may also be applied

4.5 Distance Functions

4.5.1 Numeric Attributes

Minkowski distance

dist(x 
— XJ2 Ih 
+...+lxzr 
(4) 

where h is a positive integer.

If h = 2, it is the Euclidean distance

If h = 1, it is the Manhattan distance

Weighted Euclidean distance

(L) 
). " ٦ 
٧( 
)) ٦ 
+Zlx(ZM 
١)- ٦ 
x'lx(usgp 

Squared Euclidean distance

( 8 ) 

Chebychev distance

dist(y 
) = тах(] 
Ј2 
D. 
(9) 

4.5.2 Binary and Nominal Attributes

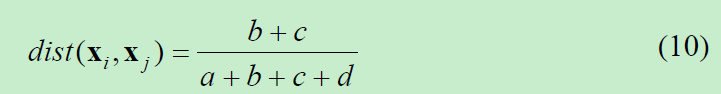
Binary Attributes

Symmetric attributes

both of its states(0 and 1) have equal importance

simple matching distance

Data point 
Data point x] 
b+d 
c+d 



We can also weight some components in Equation(10) according to application needs.

(11) 
(12) 

Asymmetric attributes

one of the states is more important or valuable than the other

Jaccard distance

dist(x X ) — 
14) 

where the number of negative matches, d, is considered unimportant and is thus ignored.To express different emphases,

dist(xpxj) 
CI c) 
(16) 

there is also a Jaccard coefficient, which measures similarity(rather than distance) and is defined as

a / (a+b+c)

Nominal Attributes

simple matching distance

Given two data points xi and xj, let the number of attributes be r, and the number of values that match in xi and xj be q:

(1)) 
) x'lx(mp 
b—.l 

4.5.3 Text Documents

1.a document can be represented as a

vector just like a normal data point

2.However, we use similarity to compare

two documents rather than distance.

3.cosine similarity function

4.6 Data Standardization

Interval-scaled attributes

There are two main approaches to standardize interval scaled attributes,range and z-score.

range

— min(f) 
rg(xzf) 
max(f) — min(f) ' 
(19) 

z-score

(x 
if 
(20) 

Where



Given the value xif, its z-score (the new value after transformation) is z(xif),

xłf — g f 
z(xłf) 
(22) 

Ratio-Scaled Attributes

Performing logarithmic transformation to each value,the attribute can be treated as an interval scaled attribute.

Nominal (Unordered Categorical) Attributes

Let the number of values of a nominal attribute be v. We can then create v binary attributes to represent them

Ordinal (Ordered Categorical) Attributes

treat ordinal attributes as interval-scaled attributes and use the same methods

4.7 Handling of Mixed Attributes

A dataset may contain any subset of the six types of attributes, interval-scaled,

symmetric binary, asymmetric binary, ratio-scaled, ordinal and nominal

attributes.

1.choose a dominant attribute type and then convert the attributes of other types to this type.(it does not make much sense to convert a nominal attribute with more than two values or an asymmetric binary attribute to an interval-scaled attribute)

logarithms of ratio variables. However, for nominal variables 
with more 
than two states this does not make much sense because some codes may be 
further apart than others without reflecting an intrinsic " remoteness" of the 
corresponding states. Also, asymmetric binary variables would be treated 
symmetrically. 

*Finding Groups in Data: An Introduction to Cluster Analysis*

*作者：Leonard Kaufman， Peter J. Rousseeuw*

2.compute the distance of each attribute of the two data points separately and then combine all the individual distances to produce an overall distance.

dist(Xl , X] ) — 
(24) 

if all δ ij ’s are 0, some default value may be used or one of the data points is removed.

dij is the distance contributed by attribute f, and it is in the range 0–1. If f is a binary or nominal attribute,

I 
0 
if Xif * x Jf 
otherwise 
(25) 

If attribute f is interval-scaled, we use



Where Rf = max( f ) − min( f )

## 读书会讨论问题汇总

1.使用k-means算法的时候，全局最小值对于大规模数据集来说在计算上是不可行的，那么有没有方法在保证计算相对较小的情况下，获得较高的相对最小值呢？

将kmeans迭代一定次（如logn）

支持向量机（核函数转换高维空间，超平面，一次聚类多次核函数）有待证实

2.空聚类是如何产生的？为什么解决空聚类的时候，选择离一个含有大量数据的聚类的聚类中心最远的数据点？

原因不详，受初始种子影响。为了减小sse

https://blog.csdn.net/shwan\_ma/article/details/80096408

3.如何判断离聚点中心最远的点是否为异常值

https://blog.csdn.net/qq\_34531825/article/details/72828182

【1】基于统计模型的方法：首先建立一个数据模型，异常是不显著属于任何簇的对象；

【2】基于邻近度的方法：通常可以在对象之间定义邻近性度量，异常对象是那些远离其他对象的对象。

【3】基于密度的方法：仅当一个点的局部密度显著低于它的大部分近邻时才将其分类为离群点。

【4】基于聚类的方法：聚类分析用于发现局部强相关的对象组，而异常检测用来发现不与其他对象强相关的对象。因此，聚类分析非常自然的可以用于离群点检测。

4.为什么层次聚类中为什么合并聚类比分裂聚类的使用要广泛？两种方法分别有什么优势和劣势，分别会在什么情况下出现？两种方法一般会分别用在什么情况下？

矩阵更新过程中，总是将两个距离最近的聚类合并，那么我们只要加入一个阈值判断，当这个距离大于阈值时，就说明不需要再合并了，此时算法结束。这样的阈值引入可以很好的控制算法结束时间，将层次截断在某一层上。

分裂可能会分错的概率大？分几份也是个问题。

5.在混合属性的处理中，为什么将一个具有多余两个状态的名词性属性或者非对称布尔属性转换为区间度量函数是没有意义的？

多于两个状态的名词性属性很难用区间度量（难以判断离谁近），不平衡的布尔值权重不均衡

6.如何保证聚类算法一定收敛呢？换言之，在算法实现时，如何判断迭代过程收敛呢？因为如果按照P100(中文版第二版)提到的三个收敛条件终止迭代，但是一个条件都无法满足，那么这时候要怎么去判断终止（收敛）条件呢？如果通过设置迭代次数终止，阈值该如何考虑？如果通过判断两次迭代结果的差别，又该从哪些角度考虑呢？

https://blog.csdn.net/u010899985/article/details/82216814

根据数学推导，k-means必收敛，至于阈值的选取和评估有待研究。

7.由于k是人为经验或随机给定的，所以难免有时候会存在空cluster的情况。但是在算法迭代计算过程中，要怎么处理空cluster的问题呢？如果再随机给定一个新质心，对最后的聚类结果肯定会有影响；如果在计算过程中改变k值，是否可行？；给定一个无穷远的点作为新的质心，貌似可行？？？

书上的做法是分裂一个已知的簇。计算过程中改变k值认为是不可行的，无穷远不可行。

8.如何找到最佳的簇数

https://blog.csdn.net/sxllllwd/article/details/82151996(手肘法，轮廓系数法)

9.K-means算法由于其本身的特性，只能在数据为球形簇的情况下具有良好的效果，那么有什么好的聚类思路可以解决非球形簇的聚类问题呢？

DBSCAN，谱聚类。聚类方法待总结