Informe 1:
Clustering

Luciano Lorenti

3 de julio de 2015

Índice general

Capítulo 1

Clustering: Conceptos Básicos

Asumamos que todos los patrones están representados en términos de características que forman vectores característicos l-dimensionales. Los pasos básicos que un experto debe seguir para desarrollar una tarea de agrupamiento son las siguientes:

1. Selección de características: Las características deben ser seleccionadas para codificar la mayor cantidad de información posible concerniente a la tarea de interés. El objetivo es obtener la mínima cantidad de información redundante.

2. Medida de proximidad: Esta medida cuantifica cuán semejantes o desemejantes son dos vectores. Es natural asegurar que todas las características seleccionadas contribuyan igualmente a la computación de la medida de proximidad y que algunas características no dominen sobre otras.

3. Criterio de clustering.

4. Algoritmo de clustering: Habiendo adoptado una medida de proximidad y un criterio de agrupamiento, este paso refiere a la elección de un algoritmo específico que revele la estructura de los agrupamientos del conjunto de datos.

5. Validación de los resultados: Una vez que los resultados de los agrupamientos han sido obtenidos, tenemos que verificar su correctitud. Esto se lleva a cabo por medio de tests.

6. Interpretación de los resultados

1.0.1. Aplicaciones del análisis de clusters

- Reducción de los datos: En muchos casos, la cantidad de datos disponibles, N, es muy grande, y como consecuencia, su procesamiento se vuelve muy demandante. El análisis de agrupamientos (cluster analysis) puede ser usado para agrupar los datos en clusters, m, ($m << N$), y procesar cada cluster como una sola entidad.

- Generación de hipótesis: se aplica análisis de agrupamientos al conjunto de datos para inferir alguna hipótesis concerniente a la naturaleza de los datos.

- Testeo de hipótesis: En este contexto, el análisis de clusters es usado para la verificación de la validez de una hipótesis específica.

- Predicción basada en grupos: Se aplica análisis de agrupamientos al conjunto de datos disponible. Los agrupamientos resultantes son caracterizados basándose en las características de los patrones que lo conforman.

1.0.2. Definición de clustering

La práctica de clasificar objetos de acuerdo a las semejanzas percibidas en la base de muchas cienciasl El cluster analysis es el estudio formal de algoritmos y métodos para agrupar o clasificar objetos. Un objeto es descripto por un conjunto de medidas or por la relación entre el objeto y otros. El cluster analysis no usa etiquetas. La ausencia de etiquetas distingue el análisis de agrupamientos del análisis de discriminante (y del reconocimiento de patrones y análisis de decisión). El objetivo del análisis de agrupamientos consiste en encontrar una organización de los datos conveniente y válida. No es el objetivo establecer reglas para separar datos futuros en categorías. Los algoritmos de agrupamientos están orientados hacia la búsqueda de estructuras en los datos.

Un agrupamiento se compone de un número de objetos similares agrupados. Everrit [**?**] [**?**] decumentó algunas de las definiciones de agrupamiento:

1. Un agrupamiento es un conjunto de entidades que son *parecidas* y entidades de diferentes agrupamientos no son parecidas.

2. Un agrupamiento es una agregación de puntos en el espacio de prueba tales que la *distancia* entre cualquiera par de puntos dentro del cluster es mas pequeña que la distancia entre cualquier punto en el cluster y cualquier otro punto fuera de el.

3. Los vectores son vistos como puntos en un espacio l-dimensional, y los clústeres son descriptos como “regiones continuas de este espacio que contienen una densidad de puntos relativamente alta, separado de otra

región de alta densidad por regiones de densidad relativamente bajas". Los clusteres definidos de esta forma se refieren como agrupamientos naturales.

Uno de los problemas cruciales en la identificación de agrupamientos en datos es especificar cuál es la función de proximidad y como medirla.

Sea X nuestro conjunto de datos $X = \{x_1, x_2, ..., x_n\}$, definimos como un m-agrupamiento de X, R, la partición de X en m conjuntos: $C_1, ..., C_m$ de forma que las tres condiciones siguientes se cumplen:

- $C_i \neq \emptyset, i = 1, ..., m$
- $\cup_{i=1}^{m} C_i = X$
- $C_i \cap C_j = \emptyset, i \neq j, i, j = 1....m$

Adicionalmente los vectores contenidos en C_i son más similares entre sí y menos similares a los vectores de los otros grupos.

La velocidad, confiabilidad y consistencia de los algoritmos de agrupamiento para organizar los datos constituyen unas excelentes razones para usarlos.

1.0.3. Represenación de los datos

Si cada objeto en un conjunto de n objetos es representado por un conjunto de d mediciones (o atributos), cada objeto es representado por un patrón o vector d-dimensional. EL conjunto mismo es visto como una matriz de patrones $n \times d$. Cada fila de la matriz define un patrón y cada columna denota una característica.

Las d características son usualmente interpretadas como un conjunto de ejes ortogonales. Los n patrones son los puntos embebidos en un espacio d-dimensional llamado espacio de patrones. Un agrupamiento puede ser visualizado como una colección de patrones que están cerca entre sí o que satisfacen alguan relación espacial. La tarea de un algoritmo de clustering es identificar estos agrupamientos naturales en espacios de varias dimensiones.

Matriz de proximidad

Los métodos de clustering requieren sea establecido un índice de proximidad o afinidad o asociación entre pares de patrones. Una matriz de proximidad $[d(i, j)]$ acumula los índices de proximidad de a pares en una matriz en la que cada columna y fila corresponden un patrón. Todas las matrices de proximidad son simétricas, de forma que todos los pares de los obejtos tienen el mismo índice de proximidad.

Una matriz de proximidad puede ser de semejanza o de desemejanza. A mayor semejanza entre el objeto i y el j, mayor será el índice de semejanza y menor el índice de desemejanza.

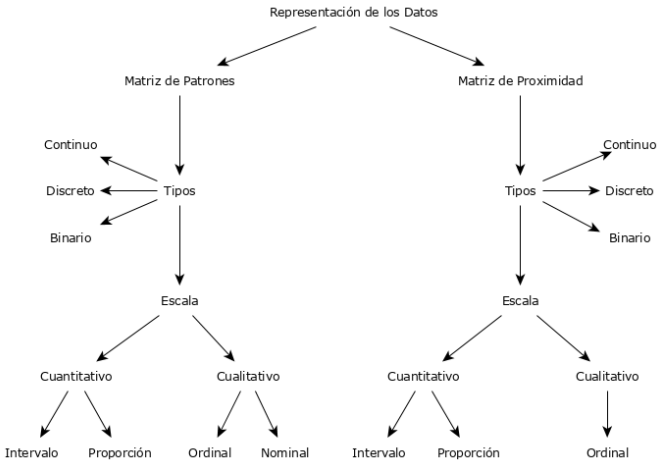

Figura 1.1: Formatos, tipos y escalas de los datos

1.0.4. Tipos de datos y escalas

Los tipos de datos refieren al grado de cuantización en los datos. Una característica puede tener varios valores de un rango continúo (subconjunto de R) o de un conjunto discreto finito. Si el conjunto discreto tiene solo dos elementos, entonces la característica es llamada binaria o dicotómica. Una característica discreta tiene un conjunto finito, usualmente pequeño, de posibles valores. A veces es conveniente pensar un valor de una característica como un punto en la recta real que puede tomar cualquier valor real en un rango fijo de valores. Estas características son llamadas continuas.

Los índices de proximidad también pueden ser discretos, binarios o continuos. Por ejemplo, supongamos que un conjunto de objetos es particionado en conjuntos mutuamente exclusivos. Un índice binario de semejanza asigna cero a los pares de objetos que se encuentran en diferentes subconjuntos y un uno a los pares que se encuentran en el mismo conjunto. Un índice de proximidad de rango (rank order) consiste en un entero entre 1 y $\frac{n(n-1)}{2}$, donde n es el número de objetos. Los enteros representan el orden relativo de las proximidades. Este índice es discreto. El índice de proximidad de distancia euclídea, definida para los patrones en un espacio de patrones, es un índice de proximidad continuo.

La segunda propiedad de una característica y un índice de proximidad es la escala de los datos. Esta propiedad indica la significancia relativa de los números. La escala de los datos pueden ser dicotomizadas en escalas cuantitativas(nominal y ordinal) y en escalas cualitativas(intervalo y proporción) Una escala nominal no es realmente una escala porque los números son simplementos usados como nombres. Una respuesta a una pregunta cuyos valores posibles son si o no puede ser codificada como $(0, 1)$ o $(1, 0)$ o $(50, 100)$. Los números no tienen signfiicado en un sentido cuantitativo. La

otra escala cualitativa es la escala ordinal, en esta los números tienen un significado solo en relación a otro. Por ejemplo $(1, 2, 3)$, $(10, 20, 30)$ y $(1, 20, 300)$ son equivalentes desde un punto de vista ordinal. Las características y los índices de proximidad binarios y discretos pueden ser codificados a escalas cualitativas.

La separación de los números tiene significado en una escala de intervalos(interval scale). Existe una medida de unidad y la interpretación de los números depende de esta unidad. La temperatura medida en grados Celsius es un ejemplo de este tipo de escalas.

Las escala más fuerte es la escala de proporción (ratio scale), en la que los números tienen un significado absoluto. Esto implica que existe un cero absoluto junto con una unidad de medición, de forma que las proporciones tienen un signfiicado. Por ejemplo las distancias medidas.

1.0.5. Medidas de proximidad

Medida de semejanza

Una medida se desemejanza es d es una función: $d : X \times X \to R$ donde R es el conjunto de números reales tal que:

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

Si en adición:

- $d(x, y) = d_0 \leftrightarrow x = y$
- $d(x, z) \leq d(x, y) + d(y, z) \forall x, y, z \in X$

d es llamada métrica.

Medida de semejanza

Una medida de semejanza, s, en X se define por $s : X \times X \to R$ tal que:

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

Si adicionalmente:

- $s(x, y) = s_0 \leftrightarrow x = y$
- $s(x, y)s(y, z) \leq [s(x, y) + s(y, z)]\, s(y, z) \forall x, y \in X$

s es llamada métrica de semejanza.

No todos los algoritmos de clustering, sin embargo, están basados en medidas de proximidad entre vectores. En los algoritmos de clustering jerárquicos uno tiene que computar distancias entre pares de conjuntos de vectores de X.

Sea U un conjunto que contiene a los vectores de X. Esto es $D_i \subset X, i = 1, ..., k, yU = \{D_1, ..., D_k\}$. Una medida de proximidad p en U es una función $P : U \times U \to R$. Se repiten las mismas ecuaciones para las medidas de semejanza y desemejanza reemplazando con D_i, D_j en lugar de x e y y U en lugar de X. Usualmente, las medidas de proximidad entre dos conjuntos D_i y D_j se definen en términos de medidas de proximidad entre elementos de D_i y D_j.

1.0.6. Medidas de proximidad entre dos puntos. Vectores continuos

Un índice de proximidad puede ser determinado de varias formas. Supongamos que comenzamos con una matriz de patrones $[x_{ij}]$, donde x_{ij} es la característica j-ésima para el patrón i-ésimo. Supongamos que todas las características son continuas y medidas en una escala de proporción. La medida o índice de proximidad mas común de semejanza entre vectores $\in R$ son la métrica pesada l_p.

El i-ésimo patrón, que corresponde a la i-ésima fila de la matriz de patrones,se denota por el vector x_i.

$$x_i = (x_{i1}, x_{i2}, ..., x_{il})^T, i = 1, ..., n$$

donde l es el número de características, n es el número de patrones. La métrica pesada l_p se define como:

$$d_p(x, y) = \left(\sum_{i=1}^{l} w_i |x_i - y_i|^p \right)^{\frac{1}{p}}$$

donde x_i, y_i son las coordenadas i-ésimas de x e y, $i = 1, ..., l$ y $w_i \geq 0$ es el i-ésimo coeficiente de peso. Si $w_i = 1, i = 1, ..., l$ obtenemos la métrica l_p no pesada o la métrica Minkowski. Si $p = 2$ tenemos la distancia euclídea:

- Norma Manhattan pesada

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

- La norma Q_∞

  $$d_\infty(x, y) = \max_{1 \leq i \leq l} |x_i - y_i|$$

La norma l_1 y la nombra l_∞ puede ser vista como una sobreestimación y una subestimación de la norma l_2 respectivamente.

$$d_\infty(x,y) \leq d_2(x,y) \leq d_1(x,y)$$

Cuando $l = 1$ todas las normas l_p coinciden.

Basado en estas medidas de desemejanza podemos definir las correspondientes medidas de semejanza como:

$$s_p(x,y) = b_{\max} - d_p(x,y)$$

Algunas medidas de semejanza son [**?**]:

- $d_G(x,y) = -log_{10}\left(1 - \frac{1}{l}\sum_{j=1}^{l}\frac{|x_j - y_j|}{b_j - a_j}\right)$

  Donde b_j y a_j son los valores máximos y mínimos entre las características J de los N vectores de X, respectivamente. En general $d_G(x,y) \neq d_G(y,x)$

- $d_Q(x,y) = \sqrt{\frac{1}{l}\sum_{j=1}^{l}\left(\frac{x_j - y_j}{x_j + y_j}\right)^2}$

La distancia de Mahalanobis al cuadrado ha sido usada como una medida de distancia en cluster analysis [**?**]. La expresión para la distancia Mahalanobis al cuadrado entre el patron x_i y el patron x_j es:

$$d(i,k) = (x_i - x_k)^T \Sigma^{-1}(x_i - x_k)$$

donde la matriz Σ es la matriz de covarianza. La distancia Mahalanobis incorpora la correlación entre las características y estandariza cada característica a media cero y varianza unitaria. Si Σ es la matriz identidad, la distancia Mahalanobis al cuadrado es la misma que la distancia euclidea al cuadrado.

El índice de correlación de muestras es un índice de semejanza para datos continuos y proporcionales que puede ser usado con patrones pero se usa más frecuentemente para medir el grado de dependencia lineal entre las dos características:

$$d(j,r) = \left|\frac{1}{n}\sum_{i=1}^{n}\frac{(x_{ij} - m_j)(x_{ir} - m_r)}{s_j s_r}\right|$$

donde m_j y s_j^2 son la media muestraly la varianza muestral. Si $d(i,j) = 0$ las características i y j son linealmente independientes.

Medidas de semejanza

- Producto interno: Se define como

$$S_{\text{inner}}(x,y) = x^T y = \sum_{i=1}^{l} x_i y_i.$$

Por lo general se usa cuando x e y están normalizados, de forma que tengan la misma longitud a. En estos casos el límite superior en inferior es $+a^2$ y $-a^2$. $S_{\text{inner}}(x, y)$ depende exclusivamente del ángulo de x e y. La medida de desemejanza correspondiente es $d_{\text{inner}}(x, y) = b_{\max} - S_{\text{inner}}(x, y)$.

- La medida de semejanza del coseno está muy relacionada con el producto interno

$$S_{\text{coseno}} = \frac{x^T y}{||x||||y||}$$

donde $||x|| = \sqrt{\sum_{i=1}^{l} x_i^2}$ e $||y|| = \sqrt{\sum_{i=1}^{l} y_i^2}$ son las longitudes de los vectores x e y respectivamente. Esta medida es invariante a las rotaciones pero no a transformaciones lineales.

- Coeficiente de correlación de Pearson: Esta medida puede ser expresada como:

$$\Gamma_{\text{pearson}}(x, y) = \frac{x_d^T y_d}{||x_d||||y_d||}$$

donde $x_d = [x_1 - \bar{x}, ..., , x_l - \bar{x}]$ e $y_d = [y_1 - \bar{y}, ..., y_l - \bar{y}]$. Con x_i, y_i siendo las coordenadas i-ésimas de x e y, respectivamente y $\bar{x} = \frac{1}{l}\sum_{i=1}^{l} x_i, \bar{y} = \frac{1}{l}\sum_{i=1}^{l} y_i$. Usualmente x_d e y_d son los llamados vectores diferencia. Γ_{pearson} toma valores entre -1 y 1. Una medida de semejanza relacionada puede ser definida como

$$D_\Gamma(x, y) = \frac{1 - \Gamma_{\text{pearson}}(x, y)}{2}$$

Esto toma valores entre $[0, 1]$. Esta medida ha sido utilizada en el análisis de genes [**?**].

- Otra medida comunmente usada es la medida de Tanimoto, también conocida como distancia como distancia Tanimoto. Se define como

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

Sumando y restando el término $x^T y$ en el denominador y luego de algunas operaciones algebráicas obtenemos

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

Es decir que la medida de Tanimoto entre x e y es inversamente proporcional a la distancia euclídea entre x e y divida por su producto interno. En el caso en el que los vectores de x han sido normalizados a la misma longitud a, la última ecuación nos conduce a:

$$S_T(x,y) = \frac{1}{1 + 2\frac{a^2}{x^T y}}$$

En este caso, S_t es inversamente proporcional a $\frac{a^2}{x^T y}$. Por lo que a mayor correlación de x e y más grande será el valor de S_t.

- Finalmente, otra medida de semejanza es la siguiente [**?**]

$$S_c(x,y) = 1 - \frac{d_2(x,y)}{||x||||y||}$$

$S_c(x,y)$ toma su máximo valor (1) cuando $x = y$ y su mínimo valor (0) cuando $x = -y$.

1.0.7. Vectores con valores discretos

Consideremos vectores x cuyas coordenadas pertenecen al conjunto finito $F = \{0, 1, ..., k-1\}$. Podemos imaginar a estos vectores como vértices en un grilla l-dimensional. Consideremos $x, y \in F^l$ y sea $A(x,y) = [a_{ij}], i, j = 0, 1, ...k-1$, una matriz $K \times K$, donde el elemento a_{ij} es el número de lugares donde el primer vector tiene el símbolo i y el elemento correspondiente del segundo vector tiene el símbolo j, $i, j \in F$. Esta matriz se denomina matriz de contingencia.

$l = 6, k = 3$ $x = [0,1,2,1,2,1]^T, y = [1,0,2,1,0,1]^T$, entonces la matriz $A(x,y)$ es igual a:

$$A(x,y) = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

$$\sum_{i=0}^{k-1}\sum_{j=0}^{k-1} a_{ij} = l$$

Medidas de desemejanza

- Distancia Hamming: Se define como el número de lugares en donde los dos vectores difieren. Usando la matriz A, podemos definir la distancia Hamming $d_H(x, y)$.

$$d_H(x, y) = \sum_{i=0}^{k-1} \sum_{\substack{j=0 \\ j \neq i}} a_{ij}$$

  Es una suma de todos los elementos fuera de la diagonal de A. En el caso especial que $k = 2$, los vectores $x \in F^l$ son binarios, la distancia Hamming se vuelve:

$$d_H(x, y) = \sum_{i=1}^{l} (x_i^2 + y_i^2 - 2x_i y_i) = \sum_{i=1}^{l} (x_i - y_i)^2$$

- la distancia l_1

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

Medidas de semejanza

- Una medida muy usada para medir la semejanza para valores discretos es la medida Tanimoto. Está inspirada en la comparación de conjuntos. Si X e Y son dos conjuntos y $n_X, n_Y, n_{X \cap Y}$ son las cardinalidades de X, Y, e $X \cup Y$ la medida de Tanimoto se define como:

$$\frac{n_{X \cap Y}}{n_X + n_Y - n_{X \cap Y}} = \frac{n_{X \cap Y}}{n_{X \cup Y}}$$

  En otras palabras, la medida de Tanimoto entre dos conjuntos es la proporción del número de elementos que tienen en común sobre el número de todos los elementos diferentes.

  Ahora nos enfocaremos en la medida de Tanimoto para dos vectores discretos x e y. La medida toma en cuenta todos los pares de coordenadas correspondientes de x e y, excepto aquellos valores cuyas coordenadas correspondientes (x_i, y_i) son 0. Definimos $n_x = \sum_{i=1}^{k-1} \sum_{j=0}^{k-1} a_{ij}$ y $n_y = \sum_{i=0}^{k-1} \sum_{j=1}^{k-1} a_{ij}$, donde a_{ij} son los elementos de $A(x, y)$. En otras palabras $n_x(n_y)$ denota el número de coordenadas no cero de $x(y)$. Entonces la medida de Tanimoto se define como:

$$S_t(x,y) = \frac{\sum_{i=1}^{k-1} a_{ii}}{n_x + n_y - \sum_{i=1}^{k-1}\sum_{j=1}^{k-1} a_{ij}}$$

Cuando $k = 2$

$$S_t(x,y) = \frac{a_{11}}{a_{11} + a_{01} + a_{10}}$$

1.0.8. Proximidad entre un punto y un conjunto

En muchas esquemas de clustering, un vector x es asignado a un cluster C teniendo en cuenta la proximidad entre x y C, $P(x, C)$. Hay dos direcciones generales para la definición de $P(x, C)$. De acuerdo a la primera, todos los puntos de C contribuyen a $P(x, C)$. Ejemplos típicos de estos casos incluyen:

1. La función de proximidad max:

$$P^{\text{ps}}_{\text{max}}(x, C) = \underset{y \in C}{max} P(x, y)$$

2. La función de proximidad min:

$$P^{\text{ps}}_{\text{min}}(x, C) = \underset{y \in C}{min} P(x, y)$$

3. La función de proximidad promedio:

$$P^{\text{ps}}_{\text{avg}}(x, C) = \frac{1}{n_c} \sum_{y \in C} P(x, y)$$

   donde n_c es la cardinalidad de C.

De acuerdo a la segunda definición; C es equipado con un elemento representativo y la proximidad entre x y C se mide como la proximidad entre x y el elemento representativo de C.

Representación en forma de punto

- El vector promedio

$$m_p = \frac{1}{n_c} \sum_{y \in C} y \text{ donde } n_c \text{ es la cardinalidad de } C$$

  Se usa por lo general para vectores contínuos.

- El centro medio $m_c \in C$ se define como

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

donde d es una medida de desemejanza entre dos puntos. Cuando se utiliza una función de semejanza la desigualda se invierte. Se utiliza, por lo general, para vectores discretos.

- El centro mediana $m_{\text{med}} \in C$ se define como el punto que

$$\text{med}(d(m_{\text{med}}, y) | y \in C) \leq \text{med}(d(z, y) | y \in C) \forall z \in C$$

donde d es la medida de semejanza entre dos puntos, med(T), con T siendo un conjunto de q escalares, es el mínimo número en T que es más grande o igual al número $\frac{q+1}{2}$-ésimo de T. Una forma algorítmica para determinar med(T) consiste en listar los elementos de T en orden ascendente y elegir el eleemento $[\frac{q+1}{2}]$.

Representación por medio de hiperplanos

Clusters con forma lineal o hiperplanar no pueden ser representados adecuadamente por un solo punto. En estos casos usamos lineas o hiperplanos como representantes de estos clusters. La ecuación general de un hiperplano H es:

$$\sum_{j=1}^{l} a_j x_j + a_0 = a^T X + a_0 = 0$$

donde $[x_1, ..., x_l]$ y $a = [a_1, ..., a_l]$ es el vector de pesos de H. La distancia de un punto x a H es:

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

Si se utiliza la distancia euclídea:

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

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

Representación hipersférica

Otro tipo de clusters son aquellos circulares o hiperesféricos en dimensiones mas altas. Para estos clusters la representación idel es por medio de un círculo (o una hipersfera). La ecuación general de una hipersfera Q es:

$$(x - c)^T (x - c) = r^2$$

donde c es el centro de la hipersfera y r su radio. La distancia entre un punto x a Q se define como:

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

1.0.9. Funciones de proximidad entre dos conjuntos

Si D_i y D_j son dos conjuntos de vectores las funciones de proximidad más coumnes son:

- La función de proximidad max:

$$P^{ss}_{\max}(D_i, D_j) = \max_{x \in D_i y \in D_j} P(x, y)$$

  Si P es una medida de desemejanza $P^{ss}_{\max}$ no es una medida. Se determiann los pares más desemejantes de vectores. Si P es una medida de semejanza $P^{ss}_{\max}$ es una medida pero no es una métrica. En este caso $P^{ss}_{\max}$ se determina por el par de vectores mas similares.

- La función de proximidad min:

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

  Si P es una medida de semejanza $P^{ss}_{\min}$ no es una medida. Se determinan por los vectores más desemejantes . Si P es una medida de desemejanza $P^{ss}_{\min}$ es una medida pero no es una métrica. En este caso $P^{ss}_{\min}$ se determina por el par de vectores mas similares.

- La función de proximidad promedio:

$$P^{ss}_{\text{avg}}(D_i, D_j) = \frac{1}{n_{D_i} n_{D_j}} \sum_{x \in D_i} \sum_{y \in D_j} P(x, y)$$

  donde n_{D_i} y n_{D_j} son las cardinalidades de D_i y D_j.

- La función de proximidad media

$$P^{ss}_{\text{mean}}(D_i, D_j) = P(m_{D_i}, m_{D_j})$$

  donde m_{D_i} es el elemento representativo de $D_i, i = 1, 2$

- Otra función de proximidad basada en la función de proximidad media, surgida como una generalización de la dada por Ward es:

$$P^{ss}_{\text{e}}(D_i, D_j) = \sqrt{\frac{n_{D_i} n_{D_j}}{n_{D_i} n_{D_j}}}$$

1.0.10. Datos Faltantes

Un problema común encontrado en la vida real son los datos faltantes. Esto significa que para algunso vectores característicos no conocemos todas sus componentes. Las siguientes son técnicas comunmente usadas para manejar esta situación:

1. Descartar todos los vectores característicos que tienen características faltantes. Se puede usar cuando el número de vectores con características faltantes es bajo comparado con el número total de vectores característicos.

2.

3. Para la característica i-ésima, encontrar la media basándose en los valores disponibles de todos los vectores característicos de X. Luego substituir este valor para los vectores donde la característica i-ésima no está disponible.

4. Para todos los pares de componentes x_i e y_i de los vectores x e y y definimos b_i como:

$$b_i = \begin{cases} 0 & \text{Si } x_i \text{ e } y_i \text{ están disponibles} \\ 1 & \text{caso contrario} \end{cases}$$

   Entonces la proximidad entre x e y se define como:

$$P(x,y) = \frac{l}{l - \sum_{i=1}^{l} b_{-i}} \sum_{\forall i: b_i = 0} Q(x_i, y_i)$$

   donde $Q(x_i, y_i)$ denota la proximidad entre dos escalares x_i e y_i. Una elección común de Q cuando se utiliza una medida de desemejanza es $|x_i - y_i|$. La idea es simple . Sea $[a, b]$ el intervalo de posibles valores de $P(x, y)$. La definición anterior nos asegura que la medida de proximidad entre xey se expanda por todo $[a, b]$, sin importar el número de características no disponibles en ambos vectores.

5. Buscar las proximidades promedio $Q_{\text{avg}}(i)$ entre todos los vectores característicos en X a lo largo de todas las componentes $i = 1...l$. Para algunos vectores X la componente i-ésima no estará. En este caso las proximidades que incluyan la característica i-ésima son excluídas de la computación de $Q_{\text{avg}}(i)$. Definimos la proximidad $\gamma(x_i, y_i)$ entre la componente i-ésima de x e y como $Q_{\text{avg}}(i)$ si al menos una de las características x_i o y_i no están disponibles.

1.0.11. Normalización

Preparar los datos para el cluster analysis requiere algún tipo de normalización que tome en cuenta la medida de proximidad. Por ejemplo, la distancia euclidea es un índice de proximidad muy común pero implicitamente asigna más peso a características con ragos más largos con respecto a aquellas de rangos mas cortos. El patrón i-ésimo de un conjunto de datos está representado por el vector columna x_i^* y el valor de la característica j-ésima para el patrón i-ésimo se denota por el valor x_{ij}^*. El asterisco denota que los datos son crudos, es decir, sin normalización. SI n es el número de patrones en el análisis, la matriz de patrones es la matriz de A de n filas y d columnas:

$$\begin{aligned} A^* &= \begin{bmatrix} x_1^* & x_2^* & ... & x_n^* \end{bmatrix}^T \\ &= \begin{bmatrix} x_{11}^* & x_{12}^* & ... & x_{1d}^* \\ x_{21}^* & x_{22}^* & ... & x_{2d}^* \\ \vdots & \vdots & ... & \vdots \\ x_{n1}^* & x_{n2}^* & ... & x_{nd}^* \end{bmatrix} \end{aligned}$$

La normalización más simple consiste en substraer la media de cada patrón:

$$x_{ij} = x_{ij}^* - m_j$$

Esta normalización hace que los valores de las características sean invariantes a desplazaientos rígidos de las coordenadas. El segundo tipo de normalización translada y escala los ejes, de forma que todas las características tengan media 0 y varianza unitaria:

$$x_{ij} = \frac{x_{ij}^* - m_j}{s_j}$$

La matriz $d \times d$, $R = [r_{ij}]$ se define en términos de los datos normalizados:

$$R = \frac{1}{n} A^T A$$

donde

$$r_{ij} = \frac{1}{n} \sum_{k=1}^{n} x_{ki} x_{kj}$$

R es la matriz de covarianza, r_{ij} el coeficiente de correlación de muestas entre la característica i y la característica j y $r_{jj} = 1 \forall j$

1.0.12. Proyecciones Lineales

Los algoritmos de proyección mapean un conjunto de patrones de dimensión d en un conjunto de dimensión m, donde $m < d$.

Una proyección lineal expresa las m nuevas características como una combinación lineal de las características originales:

$$y_i = Hx_i \quad \forall i = 1, ...n$$

donde y_i es un vector columna de dimensión m, x_i un vector de dimensión d y H es una matriz $m \times d$. Los algoritmos de proyecciones lineales son relativamente simples de usar, tienen a preservar las características de los datos y tienen propiedades matemáticas bien comprendidas.

Proyecciones por autovectores

Los autovectores de la matriz de covarianza R define una proyección lineal que reemplaza las características en los datos crudos con características no correlacionadas. Estos autovectores proveen un enlace entre el análisis de clustering y el análisis factorial. Dado que R es una matriz $d \times d$ positiva semidefinida sus autovalores son reales y pueden ser etiquetados.

$$\lambda_1 \leq \lambda_2 \leq ... \leq \lambda_d \leq 0$$

Un conjunto de autovectores correspondientes $c1, c_2, ..., c_d$ se etiquetan de la misma forma. La matriz de transformación $m \times d$, H_m se define a partir de los autovectores de la matriz de covarianza. Los autovectores son también llamados componentes principales:

$$H_m = \begin{bmatrix} \mathbf{c}_1^T \\ \mathbf{c}_2^T \\ \mathbf{c}_3^T \\ ... \\ \mathbf{c}_m^T \end{bmatrix}$$

Las filas de H_m son autovectores. La matriz proyecta los patrones en un subespacio de dimensión m cuyos ejes están en la dirección de los autovectores más grande de R:

$$y_i = H_m x_i \text{ para } i = 1, ..., n$$

Los patrones proyectados pueden ser escritos como:

$$B_m = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_n^T \end{bmatrix} = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} H_m^T = AH_m^T$$

La matriz de covarianza en el nuevo espacio

$$\frac{1}{n}B_m^T B_m = \frac{1}{n}\sum_{i=1}^{n} y_i y_i^T \tag{1.1}$$
$$= (AH_m^T)^T(AH_m^T) \tag{1.2}$$
$$= H_m A^T A H_m^T \tag{1.3}$$
$$= H_m R H_m^T \tag{1.4}$$
$$= H_m C_r^T \Lambda_r C_r H_m^T \tag{1.5}$$
$$= H_M C_r^T \Lambda_r (H_m C_r^T)^T \tag{1.6}$$

La matriz $H_m C_r^T$ se puede particionar del siguinete modo:

$$H_m C_r^T = [I|0]$$

donde I es la matriz identidad $m \times m$ y O es una matriz de ceros de $m \times (d-m)$. Entonces la matriz de covarianza en el nuevo espacio, Λ_m, se vuelve una matriz diagonal:

$$\tfrac{1}{n}B_m^T B_m = \Lambda_m = \text{diag}(\lambda_1, \lambda_2, ..., \lambda_m)$$

Esto implica que las m nuevas características obtenidas aplicando la transformación lineal definida por H_m no son correlacionadas.

La técnica para elegir un valor apropiado para m se basa en los autovalores de R. Los m autovalores de R son las varianzas muestrales en el nuevo espacio, mientras que la suma de los d autovalores es la varianza total en el espacio de patrones original. Como los autovalores están ordenados de mayor a menor, uno puede elegir un m tal que:

$$r_m = \frac{\sum_{i=1}^{m} \lambda_i}{\sum_{i=1}^{d} \lambda_i} \geq 0{,}95$$

es decir, que nos asegure que el 95 % de la varianza se retenga en el nuevo espacio. Por lo tanto una buena proyección de autovectores es aquella que retiene una gran proporción de varianza presente en el espacio de características original con solo unas pocas características en el espacio transformado.

Análisis de discriminante

El análisis de discriminante clásico intenta proyectar patrones en un espacio de menores dimensiones que el espacio original. La proyección de análisis de discriminante maximiza la dispersión entre grupo mientras que mantiene la dispersión entre grupos connstante.

Matrices de Dispersión Las matrices de dispersión juegan un rol importante en el analísis de discriminante y en el análisis de clusters. El objeto de consideración son descriptos como patrones d dimensionales que fueron separados en k grupos o k categorías.

Los patrones no normalizados en el grupo k-ésimo se denotan por los vectores columna:

$$[x_1^{*(k)}, ..., x_{n_k}^{*\ (k)}]^T$$

Sea $m^{(k)}$ la media para el conjunto k-ésimo, entonces la media conjunta está dada por:

$$m = \frac{1}{k} \sum_{i=1}^{k} n_k m^{(k)}$$

donde $n = \sum_{i=1}^{k} n_k$.

Podemos normalizar a los patrones no normalizados substrayendo la media conjunta a cada uno de ellos:

$$x_i^{(k)} = x_i^{*(k)} - m$$

Entonces la matriz de dispersión S, para la media conjunta es:

$$S = \sum_{k=1}^{k} \sum_{j=1}^{n_k} (x_j^{(k)})(x_j^{(k)})^T$$

La matriz de dispersión para el grupo k-eśimo se define por:

$$S^{(k)} = \sum_{j=1}^{n_k} (x_i^{*(k)} - m^{(k)})(x_i^{*(k)} - m^{(k)})^T$$

La matriz de dispersión intra-grupo, S_w, se define como la suma de las matrices de dispersión de cada grupo:

$$S_w = \sum_{k=1}^{k} S^{(k)}$$

Finalmente, la matriz de dispersión entre grupos, S_b,se define como la matriz de dispersión para la media de los grupos

$$\begin{aligned} S_b &= \sum_{k=1}^{k} \sum_{j=1}^{n_k} (m^{(k)} - m)(m^{(k)} - m)^T \\ &= \sum_{k=1}^{k} n_k m^{(k)} {m^{(k)}}^T - nmm^T \end{aligned}$$

Las tres matrices de dispersión están relacionadas dado que:

$$x_j^{(k)} = (m^{(k)} - m) + (x_j^{*(k)} - m^{(k)})$$

Aplicando la definición de la matriz de dispersion:

$$S = S_b + S_w$$

Es decir que la dispersión total de los datos se divide en la dispersión entre grupos y la dispersión dentro de los grupos.

Un resultado importante del análisis de discriminante es la existencia de una matriz $(k-1) \times d$ llamada H_0, con una propiedad muy intersante. Sea $x_j^{(i)}$ el patrón j-ésimo de la clase i, H_0 proyecta cada patrón en un espacio de dimensión $k-1$

$$y_j^{(i)} = H_0 x_j^{(i)},\ j = 1, ..., n_i \text{ y } i = 1, ..., K$$

y la proporción de dispersión es $\frac{|S_w|}{|S|}$ permanece constante. Esta proporción es llamada lambda estadística de Wilks. Las filas de H_0 son los autovectores correspondientes a los $k-1$ autovalores no cero de $S_w^{-1} S_B$.

1.0.13. Proyecciones no lineales

La mayor parte de los algoritmos de proyección no lineales se basan es maximizar o minimizar una función de un gran número de variables. Los algoritmos de proyección no lineales pueden ser derivados desde dos puntos de vista dependiendo de la informació a priori disponible acerca de los patrones. Si existe información de las clases de los patrones el objetivo es encontrar una proyección no lineal que reduzca la dimensionalidad y maximice la separabilidad entre las categorías. En ausencia de información de categorías, el objetivo es proyectar los patrones en un espacio de dimensión inferior reteniendo la estructura los más posible. Por ejemplo, se podría obtener una proyección que preserve las distancias entre los patrones.

1.0.14. Dimensionalidad Intrínseca

La dimensionalidad intrínseca o topológica de n patrones en un espacio d-dimensional refiere al mínimo número de parámetros libres que se necesitan para generar los patrones. La dimensionalidad intrínseca es una importante característica del conjunto de datos ya que puede determinar un número apropiado de características para representar los datos. Hay dos enfoques principales para estimar la dimensionalidad intrínseca dada una matriz de patrones. El primer acercamiento intenta aplanar el enjambre de partículas en un espacio de dimensión d. Una estimación de la dimensionalidad intrínseca puede ser obtenida calculando el número de autovalores significantes de la matriz de coavarianza de los datos, pero muchas veces este método no

funciona bien. El segundo enfoque no mueve los patrones pero estima la dimensionalidad intrínseca directamente de la información de las vecindades de los patrones.

Enfoque global

Bennet [**?**] estimó la dimensionalidad intrínseca global. Su método se basa en la observación que la varianza de la distancia entre puntos elegidos aleatoriamente en una hipersfera es inversamente proporcional a su dimensionalidad. Sea x_1 y x_2 dos vectores aleatorios independientes cuyas variables tienen una distribución uniforme dentro de una esfera de radio r en un espacio d-dimensional. La distancia euclidea normalizada entre x_1 y x_2 está dada por:

$$D = \frac{\left[\sum_{i=1}^{d}(x_{1i}-x_{2i})^2\right]^{\frac{1}{2}}}{2r}$$

La función de densidad de probabilidad de D está dada por:

$$f_D(z) = 2^d d z^{(d-1)} I_{1-z^2}(\tfrac{d+1}{2}, \tfrac{1}{2})$$

donde $I_a(p,q)$ es la función beta incompleta. Si $var(D)$ denota la varianza de D, Bennet nos muestra que

$$d \times var(D) \approx \text{constante}$$

El algoritmo de Bennet utiliza esta propiedad para aplanar el enjambre de patrones. Primero los patrones(en el espacio original) son perturbados para aumentar su varianza en las distancias entre puntos y por lo tanto reduciendo d. Luego las posiciones de los patrones se ajustan ara preservar el rank order de las distancias entre puntos en las regiones locales. Estos dos procesos son repetidos hasta que no haya un aumento significativo en la varianza de las distancias entre puntos. La dimensionalidad intrínseca está determinada por el número de atovalores de la matriz de covarianza de los datos aplanados

Enfoque local

La naturaleza global del método de los autovalores para estimar la dimensionalidad intrínseca ha conducido a la proposición de algunos métodos para mantener las propiedades locales del conjunto de datos. Estos métodos no generan una configuración de puntos o proyectan los patrones a un espacio de dimensión inferior. En su lugar estiman un número apropiado de dimensiones para representar los datos. Fukunaga y Olsen [**?**] particionan el espacio de patrones y estiman los autovalores en regionales locales en vez de computar los autovalores globles. El número de autovalores significativos de las matrices de covarianzas calculadas en cada una de las regiones son calculadas y se arma una tabla que indica el número de regiones correspondiente a cada dimensionalidad. El procedimiento es iterativo.

Capítulo 2

Métodos de clustering y algoritmos

El análisis de clusters es el proceso de clasificar objetos en subcojuntos que tienen un significado para un problema particular. Un agrupamiento es un tipo de clasficación impuesto en un conjunto finito de objetos. La matriz de proximidad es la única entrada de un algoritmo de clustering. Un arbol de problemas de clasificación sugerido por Lance y Williams [**?**]. Los nodos del arbol se explican a continuación:

1. Exclusivo versus no exclusivo. Una clasificación exclusiva es una partición del conjunto de objetos. Cada objeto pertenece a un subconjunto exactamente. Una clasificación no exclusiva o superpuesta puede asignar un objeto a varias clases. Los agrupamientos difusos son un tipo de clasificación no exclusiva en la que cada patrón es asignado a un grado de pertenencia a cada cluster en una partición.

2. Intrínseco o extrínseco. Una clasificación intrínseca usa solo la matriz de proximidad para realizar la clasficación. La clasificación intrínseca es llamada aprendizaje no supervizado en el reconocimiento de patrones debido a que no hay clases que denoten una partición a priori de los objetos. La clasificación extrínseca utiliza las etiquetas de las categorías de los objetos, junto con la matriz de proximidad. El problema es establecer una superficie discriminante que separe los objetos de acuerdo a la categoría.

Informe-1/ClasificacionClustering.png

Figura 2.1: Arbol de tipos de clasificación

3. Jerárquico vs Particional. La clasificación intrínseca y exclusivas son divididas a su vez en clasificaciones jerárquicas y particionales por el tipo de estructura impuesta en los datos. Una clasificación jerárquica es una secuencia anidada de particiones. La clasificación particional consiste en realizar una sola partición. Usaremos el término clustering para una clasifiación particional intrínseca y exclusiva y el término clustering jerárquico para una clasificación exclusiva, intrínseca y jerárquica.

Muchos algoritmos han sido propuestos para realizar clasificaciones exclusivas e intrínsecas. Las principales opciones algorítmicas utilizadas frecuentemente son:

1. Aglomerativo vs Divisivo. Un algoritmo jerárquico aglomerativo comienza ubicando a cada patrón en un cluster y gradualmente une estos clusters atómicos en clústeres más grandes. Los algoritmos de clasifiación jerárquicos divisivos invierten el procedimiento, comenzando con todos los patrones en un cluster y subdividiéndolo en partes más pequeñas. La clasificación particional puede ser caracterizada del mismo modo. Una sola partición puede ser establecida uniendo clústeres pequeños (aglomerativo) o fragmentando un clúster que contiene a todos los patrones (divisivo).

2. Serial vs simultáneo. Los procedimientos seriales trabajan con los patrones uno a uno, mientras que la clasificación simulatánea trabaja con el conjunto de patrones entero al mismo tiempo.

3. Monotética y politética. Esta opción es más aplicable en taxonomía, mientras los objetos a ser agrupados están represntados por patrones. Un algoritmo monotético de clustering usa las características una a una, mientras que, un procedimiento politético usa todas las características de una.

4. Teoría de grafos vs Algebral lineal. Podemos expresar algoritmos en términos de la teoría de grafos, utilizando propiedades como conectividad y completitud para definir clasificaciones, y expresar otros algoritmos en términos de construcciones algebráicas. La elección es por claridad, conveniencia y elección personal.

2.0.15. Clustering jerárquico

Un método de clustering jerárquico es un procedimiento para transformar una matriz de proximidad en una secuencua de particiones anidadas.

Los n objetos a ser agrupados son denotados por el conjunto H

$$H = \{x_1, x_2, ..., x_n\}$$

donde x_i es el i-ésimo objeto. Una partición C de H rompe a H en subconjuntos $\{C_1, C_2, ..., C_m\}$ satisfaciendo los siguiente:

$$C_i \cap C_j = \emptyset \quad i, j = 1....m \quad i \neq j$$
$$C_1 \cup C_2 \cup ... \cup C_m = H$$

Una partición B está anidana (nested) en una partición C si cada componente de B es un subconjunto de una componente de C. Es decir, C es formado uniendo los componentes de B. Por ejemplo, si el agrupamiento C con tres clusters y el agrupamiento Bcon singo clusters se definen de la siguiente manera, entonces B esta anidado en C. Ambos C y B son agrupamientos del conjunto de objetos $\{x_1, x_2, ..., x_10\}$.

$$C = \{(x_1, x_3, x_5, x_7), (x_2, x_4, x_6, x_8), (x_9, x_{10})\}$$
$$B = \{(x_1, x_3), (x_5, x_7), (x_2), (x_4, x_6, x_8), (x_9, x_{10})\}$$

Ni C, ni B están anidadoes en la siguiente partición, y esta partición no está anidado en C o B.

$$\{(x_1, x_2, x_3, x_4), (x_5, x_6, x_7, x_8), (x_9, x_{10})\}$$

Un clustering jerárquico es una secuenca de particiones en la que cada partición está anidada en la siguiente partición de la secuencia. Un algoritmo *aglomerativo* para el clustering jerárquico comienzo con clusters disjuntos, que ubican a cada uno de los n objetos en un cluster individual. El algoritmo de clustering dicta como la matrix de proximidad debe ser interpretada para unir dos o más de estos clústeres triviales, uniendo estos clusteres triviales en una segunda partición. El proceso es repetido para formar una secuencia de agrupamientos anidados en que el número de clusters decrece a medida que la secuencia de particiones progresa hasta alcanza un solo cluster que contiene a los n objetos, llamada cluster conjunto. Un algoritmo *divisivo* realiza la tarea en el orden inverso.

Un dendrograma es un tipo especial de estructura de arbol que provee una representación gráfica conveniente del clustering jerárquico. Un dendrograma consiste en capas de nodos, cada uno representando un cluster.. Las lineas que conectan los nodos representan los clusters que están anidados en otro. Cortar un dendorgrama horizontalmente crea un agrupamiento.

A continuación definiremos dos métodos específicos de clustering jerárquicos denominados single-link y complete-link.

Algoritmos Single-Link y Complete-Link

Comenzamos con una matriz de proximidad simétrica de dimensión $n \times n$ $D = [d(i,j)]$. Los $\frac{n(n-1)}{2}$ entradas en uno de los ladosde la diagonal principal se asumen que contiene un permutación de los enteros desde 1 a $\frac{n(n-1)}{2}$ sin que haya dos números iguales. Las proximidades están en una escala ordinal. Asumamos que las proximidades son de desemejanza: si $d(1,2) > d(1,3)$ significa que los objetos 1 y 3 son mas semejantes que los objetos 1 y 2.

Un *grafo umbral* es un grafo no dirigido de n nodos sin ciclos o múltiples aristas, en el que cada nodo representa un objeto. Un grafo umbral $G(v)$ se define para cada nivel de desemejanza v insertando una arista (i,j) entre los nodos i y j si los objetos i y j son menos desemejantes que v. Es decir:

$$(i,j) \in G(v) \leftrightarrow d(i,j) \leq v$$

Algoritmos simples para los métodos de clustering complete-link y single-link basados en grafos umbralews se detallan a continuación.

Algoritmo aglomerativo para el clustering single-link

Paso 1 Comenzar con clusteres disjuntos implicados por el grafo umbral $G(0)$, que no contiene aristas y ubica a cada elemento en un único cluster. $k = 1$

Paso 2 Formar el grafo umbral $G(k)$. Si el número de componentes (subgrafos maximales conectados) en $G(k)$ es menor que el número de clusters en el agrupamiento actual, redefinir el agrupamiento actual nombrando a cada componente conectado de $G(k)$ como un cluster.

Paso 3 Si $G(k)$ consiste en un grafo conectado, detenerse. Sino, $k = k + 1$ e ir al paso 2.

Algoritmo aglomerativo para el clustering complete-link

Paso 1 Comenzar con clusteres disjuntos implicados por el grafo umbral $G(0)$, que no contiene aristas y ubica a cada elemento en un único cluster. $k = 1$

Paso 2 Formar el grafo umbral $G(k)$. Si dos de los clústeres actuales forman un clique (subgrafo maximal completo) en $G(k)$, redefinir el clustering actual uniendo estos dos clusters en uno solo.

Paso 3 Si $k = \frac{n(n-1)}{2}$, de forma que $G(k)$ es el grafo completo en los n nodos, detenerse. Sino $k = k + 1$ e ir a 2.

Los clusteres de single-link son caracterizados como subgrafos maximales conectados, mientras que los clusteres de complete-link son cliques, o subrafos maximales completos. Alguno autores han encontrado algunas dificultades

prácticas con los cluteres formaqdos por el método single-link. Por ejemplo, los clusters de single-link facilmente se encadenan unos clusteres con otros y por lo general son desordenados. Una sola arista entre dos clústeres grandes es necesaria para unirlos. Por otro lado, los clusteres de complete-link son conservadores. Todos los pares de objetos tienen que estar relacionados antes de que formen un cluster.

Dendogramas y estructura recobrada

Un objetivo importante del análisis de los clusters jerárquicos es prover una imagen de los datos que puede ser interpretada facilmente. Cortar un dendrograma en un nivel define un agrupamiento e identifica los clusters. El nivel en si mismo no tiene significado en término de la scala de la matrix de proximidad. Un grafo de proximidad es un grafo umbral, en el que cada arista tiene un peso de acuerdo a su proximidad. El dendrograma dibujado a partir de un grafo de proximidad es llamado un dendrograma de proximidad y registra los agrupamientos y las proximidades en las que fueron formados. Un dendrograma de proximidad es dibujado una escala de proximidad a partir de una secuencia de grafos de proximidad y resalta los clusters que nacen antes y duran mucho en el dendrograma.

Los dendrogramas umbrales y de proximidad representan la estructura que el método de clustering jerárquico le impuso a los datos. Esta estructura impuesta puede ser capturada en otra matrix de proximidad llamada matrix cofenética. El acuerdo entre la matriz de proximidad y la matriz cofenética mida el grado en el cual el método de agrupamiento jerárquico captura la estructura de los datos.

Dado un agrupamiento jerárquico:

$$\{C_0, C_1, ..., C_{n-1}\}$$

conda el agrupamiento m-ésimo contiene $n - m$ clusters:

$$C_m = \{C_{m1}, C_{m2}, ..., C_{m(n-m)}\}$$

Una función de nivel, L, registra la proximidad en la que cada agrupamiento fue formado. Para un dendrograma umbral $L(k) = k$, porque los niveles en el dendrograma estan igualmente espaciado. En general

$$L(m) = \min\{d(x_i, x_j) : C_m \text{ se define } \}$$

La medida de proximidad cofenética d_C sobre los n objetos es el nivel en el cual x_i y x_j se encuentran por primera vez en el mismo cluster:

$$d_C(i, j) = L(k_{ij})$$

donde

$$k_{ij} = \min\{m : (x_i, x_j) \in C_{mq}, \text{ algún } q\}$$

La matriz de valores $[d_C(x_i, x_j)]$ es llamada matriz cofenética. Cuanto mas cerca la matriz cofenética y la matriz de proximidad dada, la jerarquía mejor encaja con los datos. No puede haber mas de $(n-1)$ niveles en el dendrograma, entonces no puede haber más de $(n-1)$ matrices cofenéticas distintas. Si se aplica los métodos complete-link y single-link a la matriz cofenética se obtienen el mismo dendrograma.

Otro algoritmos basado en teoría de grafos para single-link

Un algoritmo para clustering single-link comienza con un arbol recubridor mínimo para $G(\text{ínf})$, que es el grafo de proximidad que contiene las $\frac{n(n-1)}{2}$ aristas. A pesar que es posible derivad una jerarquía single-link a partir del arbol recubridor mínimo, el arbol recubridor mínimo no puede ser hallado a partir de un clustering jerárquico single-link.

Algoritmo de teoría de grafos para agrupamientos single-link

Paso 1 Comenzar con clusteres disjuntos, ubicando a cada elemento en un único cluster. Buscar el arbol recubridor ínimo en $G(\text{ínf})$

Repetir pasos 2 y 3 hasta que todos los objetos estén en el mismo cluster.

Paso 2 Unir los dos clusteres conectados por la arista del arbol recubridor mínimo con el menor peso para definir el siguiente cluster.

Paso 3 Reemplazar el peso de la arista seleccionada en el paso 2 por un valor mas alto que cualquier proximidad.

Un algoritmo divisivo es igual de simple. Remover las aristas del arbol recubridor mínimo siguiendo el orden de los pesos, cortando las más largas primero. Cada remoción define un nuevo agrupamiento, con aquellos objetos conectados en el arbol recubridor en cualquier etapa como pertenecientes al mismo cluster.

Algoritmos de actualización de matriz para Single-Link y Complete-Link

En esta sección discutiremos algoritmos para clustering complete-link y single-link en términos de un esquema de actualización de la matriz de proximidad. Este acercamiento fue popularizado por Johnson [**?**], quien formalizó el procedimiento. El algoritmo es un esquema aglomerativo. que borra las filas y las columnas en la matriz de proximidad a medida que los viejos clusters son unidos en nuevos.

La matriz de proximidad $n \times n$ es $D = [d(i, j)]$. Los clusteres son asignados un número de la secuencia $0, 1, ..., (n-1)$ y $L(k)$ es el nivel del k-ésimo cluster. Un cluster con número de secuencia m es denotado (m) y la proximidad entre el cluster (r) y (s) es denotado $d[(r), (s)]$.

Algoritmo de Jhonson para clustering single-link y complete-link

Paso 1 Se comienza con agrupamientos disjuntos teniendo nivel $L(0)$ y $m = 0$

Paso 2 Se buscan los pares más similares en el agrupamiento actual, digamos el par r y s, de acuerdo a:

$$d[(r),(s)] = \min\{d[(i),(j)]\}$$

donde el mínimo es sobre todos los pares de clusters en el agrupamiento actual.

Paso 3 $m = m + 1$. Unir los clústeres r y s para formar un solo cluster m. $L(m) = d[(r),(s)]$

Paso 4 Actualizar la matriz de proximidad D. borrando las filas y las columnas correspondientes a los clusters (r) y (s) y agregar una nueva fila y columna correspondiente al nuevo cluster creado. La proximidad entre el nuevo cluster denotado (r,s) y un cluster anterior (k) se define de la siguiente manera. Para el método de single-link:

$$d[(k),(r,s)] = \min\{d[(k),(r)], d[(k),(s)]\}$$

Para el método complete-link:

$$d[(k),(r,s)] = \max\{d[(k),(r)], d[(k),(s)]\}$$

Paso 5 Si todos los objetos están en el mismo cluster detenerse, sino ir a 2.

El paso 4 especifica como la matriz de desemejanza tiene que ser actualizada definiendo la formula para la desemejanza entre un nuevo cluster formado (r,s), y un cluster existente, (k) como n_k objetos. Los algoritmos singke-link y complete-link usan la desemejanza mínima y máxima respectivamente, entre los pares $\{(k),(r)\}$ y $\{(k),(s)\}$. Otros métodos de clustering puede ser definidos utilizando diferentes combinaciones de las distancias involucradas. Una formula general para el paso 4 que incluye la mayor parte de los métodos jerárquicos de clustering referenciadso está dada por:

$$\begin{aligned} d[(k),(r,s)] = {} & \alpha_r d[(k),(r)] + \alpha_s d[(k),(s)] \\ & + \beta d[(r),(s)] + \gamma |d[(k),(r)] - d[(k),(s)]| \end{aligned}$$

Método de clustering	α_r	α_s	β	γ
Single-Link	$\frac{1}{2}$	$\frac{1}{2}$	0	$-\frac{1}{2}$
Complete-Link	$\frac{1}{2}$	$\frac{1}{2}$	0	$\frac{1}{2}$
UPGMA(promedio de grupos)	$\frac{n_r}{n_r+n_s}$	$\frac{n_s}{n_r+n_s}$	0	0
WPGMA (promedio pesado)	$\frac{1}{2}$	$\frac{1}{2}$	0	0
UPGMC (centroide no pesado)	$\frac{n_r}{n_r+n_s}$	$\frac{n_s}{n_r+n_s}$	$-\frac{n_r n_s}{(n_r+n_s)^2}$	0
WPGMC (centroide pesado)	$\frac{1}{2}$	$\frac{1}{2}$	$-\frac{1}{4}$	0
Método de Ward (mínima varianza)	$\frac{n_r+n_k}{n_r+n_s+n_k}$	$\frac{n_s+n_k}{n_r+n_s+n_k}$	$\frac{-n_k}{n_r+n_s+n_k}$	0

Cuadro 2.1: Coeficientes para los algoritmos jerárquicos de actualización de matrix

Esta formula fue propuesta por [**?**]. La tabla 2.1 muestra los valores de los parámetros para los algoritmos más comunes El acrónimo PGM refiere a "método de grupos de a pares" (pair group method), el prefixo "U" y "W" refiere a no pesado y peso, respectivamente. Un método "no pesado" trata a cada objeto en un cluster de forma igual, sin importar la estructura del dendrograma. Un método pesado pesa a todos los clústeres por igual, de forma que objetos en pequeños clústeres son pesados mas fuertemente que objetos en grandes clústers. Los sufijos "A" y "C" refieren a promedio aritmético y centroides. Por lo tanto "UPGMA" significa "método de grupos de a pares no pesado que usa promedio aritmético" y "WPGMC" significa "método de grupos de a pares pesado que utiliza los centroides". Esta terminología fue utilizada por [**?**] y [**?**].

Sneath y Sokal [**?**] proveen una buena discusión de los fundamentos de estos métodos y definen otros algoritmos secuenciales aglomerativos jeráquicos y no superpuestos (SAHN). Cuando se mide la desemejanza entre un cluster existente y un cluster propsectivo, el método de single-link encuentra los pares de objetos más cercanos entre los dos clusters, el método de complete-link busca los pares mas distantes, y UPGMA y WPGMA usan el promedio aritmético de las desemejanzas. Los métodos de promedio aritmético no tienen un interpretación geométrica simple. Por otro lado los métodos UPGMC y WPGMC tienen un interpretación geométrica directa cuando los objetos están representados en un espacio d-dimensional. Los métodos de centroids miden la desemejanza entre dos clusters por medio de la distancia entre los dos centroides. El método UPGMC mide la distancia en términos del centroide calculad a partir de todos los patrones en cada cluster. El método WPGMC calcula los centroides a partir de los centroids de los dos clústeres que une para formar un nuevo cluster. El UPGMA pesa la contribución de

cada patrón igualmente, tomando en cuenta el tamaño de los clusters, mientras que WPGMC pesa los patrones en clusteres pequeños mas pesadamente que los patrones en clusters grandes.

Muchos estudios comparativos conluyen que el método de Ward [**?**], también llamado el método de mínima varianza, supera los otros métodos de clustering jerarquicos. Este método se basa en las nociones de error cuadrático popularizadas en procedimientos estadísticos como análisis de la varianza.

Supongamos que un agrupamiento fue alcanzado utilizando el método de Ward y que el siguiente agrupamiento en la jerarquía va a ser obtenido con el algoritmo de actualización de la matriz. Comenzamos con un conjunto de n patrones en un espacio d-dimensional. Sea $x_{ij}^{(k)}$ el valor de la característica j-ésima del patrón i cuando el patrón i está en el cluster k para $i = 1, .., , n_k$ y $j = 1, ..., d$. El centroide del cluster k, denotado $[m_1^{(k)}, ...m_d^{(k)}]$, es el centro del cluster, o el promedio de los n_k patrones en el cluster k:

$$m_j^{(k)} = \frac{1}{n_k} \sum_{i=1}^{n_k} x_{ij}^{(k)}$$

El error cuadrático para el cluster k es la suma de las distancias al cuadrado al centroide para todos los patrones en el cluster k.

$$e_k^2 = \sum_{i=1}^{n_k} \sum_{j=1}^{d} [x_{ij}^{(k)} - m_j^{(k)}]^2$$

El error cuadrático para todo el agrupamiento, que contiene K clusters, es la suma de los errores cuadráticos para cada uno de los clusteres individuales:

$$\Delta E_K^2 = \sum_{k=1}^{K} e_k^2$$

El método de Ward une el pair de clústeres que minimza el error cuadrático $DeltaE_{pq}^2$, el cambio en E_K^2 causado al unir el cluster p con el cluster q para formar el cluster t en el próximo agrupamiento. Dado que los errores cuadráticos para todos los clusteres excepto para los tres clusteres involucrdos permanece constante:

$$\Delta E_{pq}^2 = e_t^2 - e_p^2 - e_q^2$$

Luego de un poco de algebra, encontramos que el cambio en el error cuadrático depende solo de los centroides:

$$\Delta E_{pq}^2 = \frac{n_p n_q}{n_p + n_q} \sum_{j=1}^{d} [m_j^{(p)} m_j^{(q)}]^2$$

Los clusters p y q seleccionados para unir son los clusters que minimizan esta cantidad. El error cuadrático debe incrementarse a medida que le número de clusters decrece, pero el incremento es lo mas pequeño posible con el método de Ward. Una vez que los clusters p y q son unidos en el cluster

t, la proximidad entre todos los otros clusters y el nuevo cluster t debe ser actualizada. Sea r otro cluster distinto de p,q o t, la siguiente formula puede ser aplicada para encontrar $d[(r),(t)]$:

$$d[(r),(t)] = \frac{n_r+n_p}{n_r+n_t} d[(r),(p)] + \frac{n_r+n_q}{n_r+n_t} d[(r),(q)] - \frac{n_r}{n_r+n_t} d[(p),(q)]$$

Entrecruzamientos y monotonicidad en dendrogramas

Los algoritmos sinle-link y complete-link basados en grafos umbrales de proximidad son monotonicos, es decir, que el nivel en el que el próximo cluster se forma es siempre mayor, en uan escala de desemejanza, que el nivel del agrupamiento actual. La monotonicidad puede ser expesada en términos matemáticos haciendo referencia a la formula de actualización de la matriz. Si un método de clustering une los clusters (r) y (s) en un cluster (r,s), la monotonicidad demanda que:

$$d[(k),(r,s)] \geq d[(r),(s)]$$

para todos los clusters (k) distintos de (r) y (s). Es decir, ninguna desemejanza en la matriz actualizada puede ser menor que el elemento msa chico en la matriz anterior. Los dendrogramas generados a partir de los métodos de centroids no son monotonos y exhiben lo que se llama un entrecruzaiento.

La monotonicidad es claramente una propiedad del método de clustering y no tiene nada que ver con la matriz de proximidad. La ventaja de la formula de actualización de matriz es que la monotonicidad de cualquier algoritmo SAHN que puede ser expresado en términos de la actualización de esta formula puede ser predicho a partir de los coeficientes. Asumiendo que $\alpha_r > 0$, y $\alpha_s > 0$, Milligan (1979) proveyo el siguiente resultado:

1. Si $\alpha_r + \alpha_s + \beta \geq 1$ y $\gamma \geq 0$, el método de clustering es monótono.

2. Si $\alpha_r + \alpha_s + \beta \geq 1$ y $0 > \gamma \geq \max\{-\alpha_r, -\alpha_s\}$ el método de clustering es monótono

Métodos de clustering basados en teoría de grafos

La posibilidad de definir los algoritmos single-link y complete-link en términos de la teoría de grafos sugiere que se pueden definir otras propiedades además de la conectividad y la completitud para definir nuevos métodos de clustering. La idea es mirar la secuencia de grafos umbrales o de proximidad para detectar la aparición de un propiedad dada.

Nuevos algoritmos jerárquicos son formados cmabiando el paso 2 en la definición de los algoritmos de la sección 2.0.16. La función $Q_{p(k)}$ se define de la siguiente manera para todos los pares de clusteres $\{C_{mr}, C_{mt}\}$ en el agrupamiento $\{C_{m1}, ..., C_{m(n-m)}\}$:

$Q_{p(k)}(r,t) = \min \{d(i,j) :$ el subgrafo maximal de $G[d(i,j)]$ definido por $C_{mr} \cup C_{mt}$ es conexo y, o cumple la propiedad $p(k)$ o es completo $\}$

Siguiendo el algoritmo, los clusters C_{mp} y C_{mq} son unidos para formar el siguiente agrupamiento en la secuencia si:

$$Q_{p(k)} = \min\{Q_{p(k)(r,t)}$$

Algunos ejemplos de la propiedad p son dados a continuación:

Conectividad de nodos La conectividad de un nodo de un subrafo conexo es el número más grande n_c tal que todos los pares de noso están unidos por al menos n_c caminos sin tener nodos en común.

Conectividad de aristas La conectividad de aristas de un subgrafo conexo es el número más grande n_c tal que todos los pares de nodos están unidos por al menos n_c caminos sin tener aristas en común.

Grado de un nodo

Diámetro El diametro de un subgrafo conexo es la máxima "distancia" entre dos nodso en el subgrafo. La distancia entre dos nodos el el número de aristas en el camino más corto que los une.

Radio El radio de un subgrafo conexo es el entero más pequeño n_r tal que al menos un nodo está a una distancia n_r d etodos los nodos en el subgrafo.

Especificando una propiedad y su parámetro se define un nuevo método de clustering. Todos los clusteres deben, al menos, ser conexos. Una vez que todas las aristas han sido insertadas en el subgrafo, es completo y no se puede aplicar más propiedades

2.0.16. Clustering Particional

Las técnicas de clustering jerárquico organizan los datos en una secuencia anidada de grupos. Una característica importante de los métodos de clústering jerárquico es el impacto visual del dendrograma, que permite al analista ver como los objetos son unidos en clusteres o son divididos en niveles sucesivos de proximidad. El analista puede, entonces, intentar decidir si el dendorgrama completo describe a los datos o puede seleccionar un agrupamiento, en un nivel fijo de proximidad, que tenga sentiva para la aplicación. Nos referiremos a los métodos de clustering no jerárquicos como métodos de clustering particionales. Estos generan una sola partición de lso datos en un intento de recobrar los grupos naturales presenten en los datos. Los métodos de clustering jerárquicos requieren, por lo general, solo la matriz de proximidad de los objetos, mientras que las técnicas particionales esperan los datos en la foram de una matriz de patrones.

Las técnicas jerárquicas son populares en la ciencias biológicas y sociales porque necesitan la construcción de taxonomías. Las técnicas particionales son usadas frecuentemente en aplicaciones de ingeniería donde las particiones individuales son importantes. Los métodos de clustering particioanles son especialmente apropiados para la representación eficiente y la compresión de grandes bases de datos. Los dendrogramas son imprácticos para mas de unos pocos cientos de características.

El problema del clustering particional puede ser formulado de la siguiente forma. Dados n patrones en un espacio métrico d-dimensional, determinar una partición de los patrones en K grupos, o clusters, de forma que los patrones en un cluster sean mas semejantes entre sí comparado con los patrones en diferentes clusters. El valor de K puede o no estar especificado. Un criterio de agrupamiento, como el error cuadrático, debe ser adoptado. Los criterios pueden ser clasificados como gloabales o locales. Un criterio global representa cada cluster con un prototipo y asigna los patrones a los clusteres de acuerdo a los prototipos mas simiares. Un criterio local forma los clusters utilizando la estrutura local en los datos. Por ejemplo, los clusteres pueden ser formados identificando regiones de alta densidad en el espacio de patrones o asignando un patrón a sus k vecinos mas próximos al miso cluster.

La solución teórica al problema particional, es sencialla. Simplemente se selecciona un criterio, se evalúa para todas las posibles particiones que contiene k clusters, y se elige la partición que optimiza el criterio. La primera dificultad encontrar es como seleccionar un criterio que traduzca las nociones intuitivas acerca de un cluster en una fórmula matemática razonable. La segunda dificultad con este enfoque es que el número de particiones es astronómico, incluso para una cantidad de patrones moderada, de forma que evaluar el criterio más simple para todas las particiones es impráctico.

Sea $S(n,k)$ el número de agrupamientos de n objetos en K clusters. El orden de los objetos en cada cluster y el orden de los clusters no tienen importancia. Clusteres vacíos no son contados. Fortier y solomon [**?**] y Jensen [**?**] mostraros que:

$$S(n,K) = \frac{1}{K!}\sum_{i=1}^{K}(-1)^{K-i}\binom{K}{l}(i)^n$$

Estos son los números de Stirling de segundo tipo.

Claramente, la numeración exhaustiva de todas las posibles particiones no es computacionalmetne viable incluso para un pequeño número de patrones. Para evitar esta explosión de combinaciones, la función criterio es evaluada solo para un prqueño grupo de particiones "razonables". La forma de identificar este pequeño conjunto es optimizar la función criterio usando una técnica iterativa. Se comienza con una partiicon inicial, los objetos son movidos desde un cluster a otro con el objetivo de mejorar el valor de la función criterio. Por lo tanto, cada partición sucesiva es un perturbación de

la anterior, y por lo tanto, solo un pequeño numero de particiones es examinada. Los algoritmos basados en ésta técnica son computacionalmente eficientes pero tienen a converger en mínimos locales de la función criterio. Existen mucha heurísticas para elegir la partición inicial, moviendo los objetos de un cluster o a otro, y para unir y dividir grupos.

No existe un criterio "mejor" para obtener una partición porque no hay una definición precisa de lo que es un cluster. Los clusteres pueden tener formas y tamaños arbitrarios en un espacio de patrones multidimensional. Cada criterio de clustering impone una cierta estructura a los dastos, y si suecede que los datos cumplen con los requerimientos de un criterio particular, entonces los verdaderos cluteres son recobrados.

Criterio de clustering de error cuadrático

La estrategia de partición mas comun está baasada en el criterio del error cuadrático. El objetivo general es obtener la partición tal que, para un número fijo de clusteres, minimice el error cuadrático. Minimizar el error cuadrático, o la variación intra-cluster, es equivalente a maximizar la variación entre los clusteres.

Supongamos que el conjunto de n patrones dado en d dimensiones ha sido particionado de algún modo en K clusters $\{C_1, C_2, ..., C_K\}$ de forma que el cluster C_K tiene n_K patrones y cada patrón está en exactamente un cluster. El error cuadrático para el cluster C_K es la suma de las distancias euclideas al cuadrado de cada patrón en C_K y el centroide del cluster $m^{(k)}$. Este error cuadrático es también llamado la variación intra-cluster:

$$e_k^2 = \sum_{i=1}^{n_k} (x_i^{(k)} - m^{(k)})^T (x_i^{(k)} - m^{(k)})$$

El error cuadrático para el agrupamiento entero que contiene a los K clusters es la suma de las variaciones intra-clusters:

$$E_k^2 = \sum_{k=1}^{K} e_k^2$$

El objetivo del método de clustering del error cuadrático es encontrar una partición que contenga K clusters que minimice E_k^2 para un K fijo. La partición resultante también se la conoce como partición de mínima varianza

Gordon y Hedenrson [**?**] definieron el problema de clustering en términos de la minimización de la suma de las distancias al cuadrado intra-cluster. Sin embargo, escribieron su función criterio de forma tal que el problema de clustering puede ser formulado como un problema de programación no lineal. Sea x_{ij} la característica j-ésima del patrón i-ésimo, $i = 1, ..., m; j = 1, ..., d$ Sea $y_{ik} =!$ si el patrón i-ésimo pertenece al cluster k-ésimo, y 0 si el patrón i-ésimo no pertenece al cluster k-ésimo, $k = 1, ..., K$. El centroide del cluster k-ésimo, $z_k = (z_{k1}, z_{k2}, ..., z_{kd})$, donde:

$$z_{kj} = \frac{\sum_{i=1}^{n}(y_{ik}x_{ij})}{\sum_{i=1}^{n} y_{ik}}$$

La variación total intra-cluster, denotada S_T, puede ser escrita como:

$$S_T = \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik}\sum_{j=1}^{d}(x_{ij} - z_{kj})^2$$

Gordon y Henderson propusieron la siguiente formulación para minimizar S_T:

1. Dada la matriz de patrones, H, y el número de clusters, K, encontrar la matrix $Y \in 0, 1^{n\times K}$, $Y = [y_{ik}]$ tal que minimiza S_T. Y tiene un 1 en cada fila y al menos un uno en cada columna.

2. Una formulación mas general minimiza S_T bajo la suposición que $y_{ik} \in [0, 1]$ sujeto a la restricción

$$\sum_{k=1}^{K} y_{ik} = 1 \text{ y } y_{ik} \geq 0$$

   El término y_{ik} denota la fracción del patrón i-ésimo que se asignada al cluster k-ésimo.

Métodos de clustering de error cuadrático

La idea básica de un altoritmo de clustering iterativo es empezar con una partición inicial y asignar patrones a clusteres de forma de reducir el error cuadrático. El error cuadrático tiene a disminuir a medida que el número de clusteres aumenta y puede ser minimizado para un número fijo de clusteres.

Algoritmo para clustering particional iterativo

Paso 1 Seleccionar una partición inicial con K clusters

Repetir pasos 2 a 5 hasta que se estabilicen las pertenencias a los clusters

Paso 2 Generar una nueva partición asignando cada patrón a su centro de cluter más cercano.

Paso 3 Calcular los nuevos centros de los clusters como centroides de los clusteres

Paso 4 Repetir paso 2 y 3 hasta que se alcance el valor óptimo de la función criterio.

Paso 5 Ajusta el número de clústers uniendo y dividiendo los clusteres existentes o eliminando clusteres pequeños o outliers.

Partición inicial . Una partición inicial puede ser formada primeramente especificando un conjunto de K puntos iniciales. Los puntos iniciales pueden ser los primeros K patrones o K patrones elegidos aleatoriamente a partir de la matriz de patrones. Un conjunto de K patrones que se encuentra bien separados entre sí puede obtenerse tomando el centroide de los datos como el primer punto inicial, y seleccionar los puntos sucesivos que estén a no menos de una cierta distancia del punto elegido. La partición inicial se forma asignando cada patrones al punto inicial mas cercano. Los centroides de los clusteres resultantes son los centros iniciales de los clusteres. También se puede usar clustering jerárquico para seleccionar una partición inicial.

Diferentes particiones iniciales puede conducir a diferentes agruapamientos finales porque los algoritmos basados en el error cuadrático puede converger a mínimos locales. Esto ocurre especialmente si los clusteres no están bien separados. Una forma de sobreponerse al mínimo local es ejecutar el agloritmo particional con muchas particiones diferentes. Si todas conducen a la misma partición final, podemos tener cierta confianza que se alcanzo el mínimo global.

Actualización de la partición . Las particiones son actualizadas reasignando los patrones a clusteres en un intento de reducir el error cuadrático. El término pasada o ciclo refiere al proceso de examinar las etiquetas de los clusters de cada patrón una vez. McQueen [**?**] define una pasada K-medias como la asignación de todos los patrones al centro de cluster mas cercano. El centro del cluster nuevo es recomputado luego de cada asignación en el método de McQueen. El método de Forgy (1965) recalcula los centros de cada cluster luego de que fueron examinados todos los patrones

Friedman and Rubin [**?**] define una pasada hill-climbing y una pasada forcing (forcing pass) en su algoritmo de clustering. La pasada hill-climibing cambia las etiquetas ed los clusters de un patrón solo para mejorar la función criterio. La pasada k-medias asigna cada patrón a su centro de cluster mas cercano. Una pasada forcing perturba la partición par aevitar quedar atrapado en un mínimo local. La pasada forcing intenta ubicar a cada patrón de un un cluster en otro cluster. La función criterio es recalculada luega de cada test y la mejor partición es retenida, y la pasada forcing se repite para el siguiente cluster. Estas pasadas se aplican repetidamente hasta que se obtiene la covnergencia.

Ajustanto el número de clusters Algunos algoritmos de clustering pueden crear nuevos clusters o unir clusteres existentes si ciertas condiciones se cumplen. Esta capacidad le permite a un aglrotimo recobrarse desde particiones iniciales deficientes y le permite seleccionar un número de cluster natural o adecuado. En el algoritmo ISODATA, un cluster es dividido si tiene mucha patrones y una varianza muy grande en la característica mas

dispersa. Dos clusteres son unidos si los centros de los clusteres están lo suficientemente cerca, nuevamente bajo un parámetro provisto por el usuario.

Un outlier es un patrón que está lo suficientemente alejado del resto de los datos sugiriendo que fue incluido por error. Muchas veces un outlier se debe a ruido en el proceso de medición. Los outliers puede proveer información util acerca de la generación de datos subyacente, pero forzar a que un outlier pertenezca un un cluster distorciona su forma.

Convergencia. Los algoritmos particiones terminan cuando la función criterio no puede ser mejorada. No hay garantía que un algoritmo iterativo alcance el mínimo global. Algunos algoritmos se detienen cuando las etiquetas de los clusteres no varían entre dos iteraciones sucesivas. Un máximo número de iteraciones puede ser espeficiado para prevenir oscilaciones sin fin.

Programas para clutering de error cuadrático medio

FORGY es el método mas simple, solo utiliza la pasada k-medias. Los centros de los clusters son actualizados recalculando los centroids de todos los patrones que tienen las mismas etiquetas al final de la pasada. Los puntos iniciales son K patrones elegidos aleatoriamente, donde K es especificado por el usuario. La implementación de Jain y Dubes le permite al usuario especificar una heurística que crea clusteres adicionales. Luego de que el error cuadrático converge para un K fijo, se crea un nuevo cluster cuando se encuentra un patrón que está lo suficientemente lejos de los clusteres existentes. La distancia promedio entre el patrón x_i y el centro del cluster K es:

$$d_i = \frac{1}{K} \sum_{k=1}^{K} d(x_i, m^{(k)})$$

Un nuevo cluster es creado centrado en el patron x_i si

$$|d(x_i, m^{(q)}) - d_i| \leq d_i T_1$$

donde el centro q-ésimo cluster es el centro del cluster mas cercano al patrón x_i y T_1 es un parámetro especificado por el usaurio, $0 < T_1 < 1$. La parte izquierda de la desigualdad es d_i para patrones cercanos a un cluster existente y es pequeña para patrones alejados de todos los clusteres existentes. Un valo de T_1, mas grande, mayor cantidad de nuevos clusteres se crearán. FORGY también detecta outliers. Si el número de patrones en cualquier cluster cae por debajo de otro parámetro especificado por el usuario, T_2, entonces todos los patrones pertenecientes a ese cluster son considerados outliers y son ignorados.

El segundo algoritmo de clustering se llamado CLUSTER [**?**] genera un secuencia no jerárquica de agrupamientos. El programa utiliza una pasada

k-medias y una pasada forcing. CLUSTER intente encontrar los mejores agruapmientos conteniendo, 1, 2, ..., K clusters.

Involucra dos fases que son repetidas hasta que una pasada de las dos fases no decrece el error cuadrático medio. La fase 1 crea un secuencia de clusteres conteniendo $2, 3, ..., K$ clusters, donde K es especificado por el usuario. Los dos centros de los clusteres iniciales son lso centroides de los patrones y el patrón más alejado del centroide, sin contar los outliers. Dada un agrupamiento con k clusters, el patrón más alejado de los clústeres existentes se identifica como el $k + 1$-ésimo centro de cluster. La pasada k-medias se repite hasta que no hay más cambios de clusters o hasta que se alcanza un número máximo de iteraciones.

La primera pasada a través de la fase 1 da un conjunto de K agrupamientos, cada uno conteniendo un número diferente de clusters. La fase 2 crea otro conjunto de agrupamientos uniendo clusteres existentes dos por vez, para ver si se puede obtener un mejor agrupamiento puede ser obtenido. Luego de cada pasada por las fases, el error cuadrático de los agrupamientos son comparados con los agruapamientos previos. Si en alguno de los errores son mas pequeños que antes, se hace otra pasada por las fases 1 y 2. Esto continúa hasta que el error cuadrático no decrezca.

En conclusión, los programas de clustering que minimizan el error cuadrático son muy prácticos. Intentan definir clusteres que tienen una forma hyperelipsoidal.

Clustering mediante mixture decomposition

La distribución de mixturas gaussianas puede ser escrita como una superposición lineal de gaussianas de la forma

$$p(x) = \sum_{k=1}^{k} \pi_k N(x|\mu_k, \Sigma_k)$$

Introduzcamos una variable aleatoria k-dimensional z que tiene 1 de k representaciones en las que un elemento particular z_k es igual a 1 y todos los otros elementos son iguales a 0. Los valores de k satisfacen: $z_k \in \{0, 1\}$ y $\sum_k z_k = 1$, y vemos que el vector z tiene k posibles estados de acuerdo a que elemento no se encuentra en cero. La distribución marginal sobre z es especificada en términos de los coeficientes de mixtura π_k tal que:

$$p(z_k = 1) = \pi_k$$

Donde el parámetro $\{\pi_k\}$ debe satisfacer:

$$0 \leq \pi_k \leq 1$$

$$\sum_{k=1}^{K} \pi_k = 1$$

Para que sean probabilidades válidas. Debido a que z usa 1 de las K representaciones, podemos escribir esta distribución de la forma:

$$p(z) = \prod_{k=1}^{K} \pi_k^{z_k}$$

Similarmente la distribución condicional de x dado un valor particular de z es una gaussiana:

$$P(x|z_k = 1) = N(x|\mu_k, \Sigma_k)$$

que a su vez puede ser escrita como:

$$P(x|z) = \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k}$$

La distribución conjunta $p(x, z)$ está dada por $p(z)p(x|z)$ y la distribución marginal de x se obtiene sumando las distribuciones conjunta sobre todos los posibles estados de z:

$$\begin{aligned} p(x) &= \sum_z p(z)p(x|z) \\ &= \sum_z \prod_{k=1}^{K} \pi_k^{z_k} \prod_{k=1}^{K} N(x|\mu_k, \Sigma_k)^{z_k} \\ &= \sum_z \prod_{k=1}^{K} \pi_k^{z_k} N(x|\mu_k, \Sigma_k)^{z_k} \end{aligned}$$

Como solo uno ed los elementos del vector z es 1, mientras qeu todos los demás son ceros, esta expresión puede ser reescrita como

$$= \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)$$

Por lo tanto, al distribución marginal de x es una mixtura gaussiana. Si tenemos varias observaciones $x_1, x_2, ..., x_n$, entonces, debido a que hemos representado la distribución marginal en la forma $p(x) = p(z)p(x, z)$, se deduce que para cada punto observado x_n existe una variable latente correspondiene z_n. Otra cantidad que juega un rol importante es la probablidad condicional de z dado x. Usaremos:

$$\begin{aligned}\gamma(z_k) &= p(z_k = 1|x) \\ &= \frac{p(z_k = 1)p(x|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x|x_j = 1)} \\ &= \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|u_j, \Sigma_j)}\end{aligned}$$

Podemos ver a π_k como la probabilidad a priori de $z_k = 1$, y la cantidad $\gamma(z_k)$ como la probabilidad a posteriori una vez que ha sido observado x. También puede ser visto como la responsabilidad del componente k para "explicar" la observación x.

Máxima verosimilitud Supongamos que tenemos un conjunto de datos observados $\{x_1, ..., x_N\}$, y queremos modelar los datos usando una mixtura de gaussianas. Podemos representa rel conjunto de datos como una matriz $N \times D$, denominada X, en el que la n-ésima fila está dada por x_n^T. Similarmente, las variables latentes correspondientes serán denotadas por una matriz $N \times K$ denominada Z con filas z_n^T. Teniendo en cuentra que:

$$\begin{aligned}p(x) &= \sum_z p(z)p(x|z) \\ &= \sum_{k=1}^{K} \pi_k N(x|\mu_k, \Sigma_k)\end{aligned}$$

La funciónde log likelihood está dada por:

$$ln(p(x|\pi, \mu, \Sigma)) = \sum_{n=1}^{N} ln \left\{ \sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k) \right\}$$

Es importante destacar que hay un problema muy importante asociado a la maximización del framework de la verosimilitud aplicado al modelo de mixturas gaussianas, debido a la presencia de singularidades. Adicionalmente, una mixtura de k componentes tendrá un total de $k!$ soluciones equivalentes que se corresponden con las $k!$ factorial formas de asignasr k conjuntos de parámetros a los k componentes. Es decir que para cualquier punto dado en el espacio de parámetros hay $k! - 1$ puntos adicionales que conducen a la misma distribución. Este problema se conoce como identificabilidad y es una cuestión importante cuando queremos interpretar los parámetros descubiertos por el modelo.

Maximizar la función log likelihood para un modelo de mixturas gaussianas resulta un problema más complejo que para el caso de una gaussiana simple. La dificultad surje por la presencia de la sumatoria dentro del logaritmo, de forma que la función logaritmo no actúa sobre la gaussiana. Si ponemos la derivada de la log likelihood a cero, no btendremos una solución cerrada.

Expectation Maximization para mixturas gaussianas Un método elegante y poderoso para encontrar las soluciones de la máxima verosimilitud para modelos con variables latentes se llama el algoritmo de expectation-maximization)[**?**][**?**].

Comenzaremos escribiendo las condiciones que deben satisfacer el máximo de la función de verosimilutyd. Igualando las derivadas de $\ln\{x|\pi,\mu,\Sigma)\}$ con respecto a las medidas de u_k de las gaussianas a cero obtenemos:

$$0 = -\sum_{n=1}^{N} \underbrace{\frac{\pi_k N(x_n|\mu_k,\Sigma_k)}{\sum_j \pi_j N(x_n|\mu_j,\Sigma_j)}}_{\gamma(z_{nk})} \sum_k (x_n - \mu_k)$$

Multiplicando por Σ_k^{-1} y reacomodando obtenemos

$$\mu_k = \frac{1}{N^k}\sum_{n=1}^{N}\gamma(z_{nk})x_n$$

donde

$$N_k = \sum_{n=1}^{N}\gamma(z_{nk})$$

Podemos interpretar a N_k como el número de puntos asignados al cluter k. La medida u_k para la k-ésima gaussiana se obtiene calculando la media pesada de todos los puntos del conjunto de datos, en donde el factor de peso para el punto x_n está dado por la probabilidad a posteriori$\gamma(z_{nk})$ de que la componente k fue responsable de generar x_n.

Si igualamos la derivada de $\ln\{p(x|\pi,\mu,\Sigma)\}$ con respecto a Σ_k a cero obtenemos:

$$\Sigma_k = \tfrac{1}{N_k}\sum_{n=1}^{N}\gamma(z_{nk})(x_n-\mu_k)(x_n-\mu_k)^T$$

Finalmente maximizamos $\ln\{p(x|\pi,\mu,\Sigma)\}$ con respecto a los coeficientes de mixtura π_k. Debemos tener en cuenta la restricción $\sum_{k=1}^{K}\pi_k = 1$. Esto puede ser alcanzado utilizando un multiplicador de Lagrange y maximizando al siguiente cantidad:

$$\ln\{p(x|\pi,\mu,\Sigma)\} + \lambda(\sum_{k=1}^{K}\pi_k - 1)$$

lo que nos da:

$$\begin{aligned} g(\pi_k, x_n, \mu_k, \Sigma_k) &= \ln\{p(x|\pi,\mu,\Sigma)\} + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) \\ &= \sum_{n=1}^{N} ln\left[\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)\right] + \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right) \end{aligned}$$

Derivando

$$\frac{\partial g}{\partial \pi_k} = \sum_{n=1}^{N} \frac{1}{\sum\limits_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)} N(x_n|\mu_k,\Sigma_k) + \lambda$$

Igualando a 0

$$0 = \sum_{n=1}^{N} \frac{1}{\sum\limits_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)} N(x_n|\mu_k,\Sigma_k) + \lambda$$

Multiplicando por a ambos miembros por $\sum\limits_{k=1}^{K}\pi_k$

$$\begin{aligned} 0 &= \sum_{k=1}^{K}\pi_k \sum_{n=1}^{N} \frac{1}{\sum\limits_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)} N(x_n|\mu_k,\Sigma_k) + \lambda \\ 0 &= \sum_{n=1}^{N} 1 + \lambda \\ \lambda &= -N \end{aligned}$$

Reemplazando λ

$$= \sum_{n=1}^{N} ln\left[\sum_{k=1}^{K}\pi_k N(x_n|\mu_k,\Sigma_k)\right] - N\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

Derivando

$$\frac{\partial g}{\partial \pi_k} = \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)} N(x_n|\mu_k, \Sigma_k) - N$$

$$N = \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)} N(x_n|\mu_k, \Sigma_k)$$

Multiplicando por π_k a ambos miembros

$$\pi_k N = \pi_k \sum_{n=1}^{N} \frac{1}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)} N(x_n|\mu_k, \Sigma_k)$$

$$\pi_k N = \sum_{n=1}^{N} \frac{\pi_k N(x_n|\mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k N(x_n|\mu_k, \Sigma_k)}$$

$$\pi_k N = \sum_{n=1}^{N} \gamma(z_{nk})$$

$$\pi_k = \frac{N_k}{N}$$

Por lo tanto, el coeficiente de mixtura para el k-ésimo componente está dado por la responsabilidad promedio que ese componente tiene para explicar los patrones.

Estos resultados, si bien no constituyen una forma cerrada, sugieren un esquema iterativo para encontrar una solución al problema de máxima verosimilitud. Primero se eligen algunos valores iniciales para las medias, covarianzas y coeficientes de mixtura. Luego se alternan entre dos etapas de actualización: En la etapa E, se usan los valores actuales de los parámetros para evaluar las probabilidades a posteriori, o responsabilidades, dado por:

$$\gamma(z_k) = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$$

Luego usamos estas probabilidades en la etapa de maximización para reestimar las medias, covarianzas y coeficientes de mixtura. Cada actualización de los parámetros que resulta de aplicar el paso E seguido del M garantiza aumentar la función de log likelihood. En la práctica se considera que el algoritmo ha convergido cuando el cambio en la función de log likelihood, o alternativamente en los parámetros, cae por debajo de un umbral.

EM para mixturas gaussianas Dado un modelo de mixturas Gaussianas, el objetivo es maximizar la función de log likelihood con respecto a los parámetros (las medias, las covarianzas de los componentes y los coeficientes de mixtura).

1. Inicializar las medias u_k, las covarianzas Σ_k y los coeficientes de mixtura pi_k y evaluar los valores iniciales de las funciones de log likelihood.

2. Paso E. Evaluar las responsabilidades usando los valores actuales de los parámetros.

$$\gamma(z_k) = \frac{\pi_k N(x|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j N(x|\mu_j, \Sigma_j)}$$

3. Paso M. Re estimar los parámetros usando las responsabilidades actuales.

   - $\mu_k^{\text{nuevo}} = \frac{1}{N^k} \sum_{n=1}^{N} \gamma(z_{nk}) x_n$
   - $\Sigma_k^{\text{nuevo}} = \frac{1}{N_k} \sum_{n=1}^{N} \gamma(z_{nk})(x_n - \mu_k^{\text{nuevo}})(x_n - \mu_k^{\text{nuevo}})^T$
   - $\pi_k^{\text{nuevo}} = \frac{N_k}{N}$

   donde $N_k = \sum_{n=1}^{N} \gamma(z_{nk})$

4. Evaluar la función de log likelihood.

   $$ln\left[p(X|\mu, \Sigma, \pi)\right] = \sum_{n=1}^{N} ln\left[\sum_{k=1}^{K} \pi_k N\left(x_n|\mu_k, \Sigma_k\right)\right]$$

   y chequear la convergencia de algunos de los parámetros de la función de log likelihood. Si el criterio de Convergencia no se satisface, volver al paso 2.

Clustering mediante busca de moda

Los clusteres pueden ser vistos como regiones del espacio de patrones en el que los patrones son densos, separado por regiones de baja densidad de patrones. Los clusteres pueden ser identificados mediante la búsqueda de regiones de alta densidad, llamadas modas, en el espacio de patrones. Cada moda se asocia con un centro de cluster y cada patrón se asigna al cluster con el centro mas cercano. La densidad de probabilidad estimada en un punto x es proporcional al número de patrones, k_n, que caen dentro de una pequeña región de volumen V_n alrededor de x (Duda-Hart, 1973, Silvernman, 1986)

$$\hat{p}_n(x) = \frac{k_n/n}{V_n}$$

donde n es el número total de patrones. Para un V_n, k_n fijo, sera grande para puntos que yacen en una región densa, resultando en un gran estimación de $\hat{p}_n(x)$. La elección de V_n es crítica cuando n es pequeño y es gobernada ya acer por un enfoque de ventanas Parzen y por un enfoque de vecinos mas próximo.

El volumen V_n en el enfoque de la ventana Parzen se espeficia como una función de n. En el enfoque de los vecinos más próximos, k, es especifica como una función de n. La región alrededor de cada patrón es examinada para capturar sus k_n vecinos más próximos. En ambos enfoques, se argumenta que V_n es inversamente proporcional a $\sqrt{n}$. La principal diferencia entre estos dos enfoques es que la ventana alrededor de cada punto en el enfoque de la ventana Parzen tiene el mismo volumen, mientras que el tamaño de la ventana depende de la ubicación del patrón en el espacio de patrones en el enfoque de vecinos más próximos.

Mean Shift [?]

DBSCAN [?] La idea clave es que para cada punto de un cluster, la vecindad de un radio dado tiene que contener como mínimo un número de puntos, es decir, la densidad dn la vecindad tiene que exceder un umbral. El tamaño de la vecindad es determinado por la elección de una función de distancia para dos puntos p y q, denotado $dist(p, q)$.

Definición 1: La vecindad-ϵ La vecindad-ϵ de un punto p, denotado por $N_\epsilon(p)$, se define por $N_\epsilon(p) = \{q \in D | dist(p, q) \leq eps\}$ Una enfoque naif puede requerir para cada punto en el cluster la existencia de al menos un mínimo número de puntos (MinPts) en una vecindad-ϵ de ese punto. Sin embargo, este enfoque falla porque hay dos clases de puntos en un cluster, los puntos que yacen dentro del cluster (puntos core) y puntos en el borde del cluster (puntos de borde). En general, una vecindad-ϵ de un punto de borde contiene menos puntos que en la vecindad-ϵ de un punto central. Por lo tanto, tenemos que establecer el número mínimo de puntos a un valor relativamente bajo con el objetivo de incluir todos los puntos pertenecientes al mismo cluster. Este valor, sin embargo no será característico para el cluster respectivo. Por lo tanto, necesitamos que para cada punto p en un cluster C haya un punto q en C de forma que p esté dentro de la vecindad-ϵ de q y $N_\epsilon(q)$ contenga al menos MinPts puntos.

Definición 2: Directamente Densamente alcanzables Un punto p es directamente densamente alcanzable desde un punto q con respecto a ϵ y MinPts si

1. $p \in N_\epsilon(q)$
2. $|N_\epsilon(q)| \geq MinPts$

La alcanzabilidad densa directa es simétrica para puntos dentro del cluster. Pero no será simétrica, por lo general, para un punto del centro y uno del borde.

Definición 3: Densamente alcanzable Un punto p es densamente alcanzable desde un punto q con respecto a ϵ y MinPts si existe una cadena de puntos $p_1, ..., p_n, p_1 = q, p_n = p$ tal que p_{i+1} es directamente densamente alcanzable desde p_i.

Definición 4 Un punto p es conectado por densidad a un punto q con respecto a ϵ y $MinPts$ si hay un punto o de forma que p y q son densamente alcanzables desde o con respecto a ϵ y $MinPts$

Definición de cluster basándose en densidad Un cluster se define como un conjunto de puntos densamente conectados que es maximal con respecto a la alcanzabilidad por densidad. El ruido será definido relativo a un conjunto de clusters. El ruido es simplemente el conjunto de pountos en D que no pertenece a ninguno de los clusteres.

Definición 5 (cluster) Sea D una base de datos de puntos. Un cluster C con respecto a ϵ y $MinPts$ es un subconjunto no vacío de D que satisface las siguientes condiciones

1. $\forall p, q$: si $p \in C$ y q es densamente alcanzable desde p con respecto a ϵ y $MinPts$, entonces $q \in C$ (Maximal)
2. $\forall p, q, \in C$: p es densamente conectado a q con respecto a ϵ y $MinPts$(conectividad)

Dado los parámetros ϵ y $MinPts$ podemos descubrir un cluster en unb enfoque de dos pasos. Primero elegir un punto arbitrario de la base de datos que satisfaga las condiciones de un punto centrl como punto inicial. Luego, obtener todos los puntos que son densamente alcanzables desde el punto inicial para obtener los clusteres que contienen la semilla.

Clustering por teoría de grafos

1. Construir el MST para el conjunto de patrones dado
2. Identificar aristas inconsistente en el MST
3. Remover las aristas incosistentes para formar componentes conectados y llamarlos clusters

El paso crucial es definir que es una arista incosistente. Una arista es incosistente si su peso (distancia entre dos puntos) es signfiicativamente más grande que el promedio de las aristas cercanas. Por lo que la arista inconsistente se relaciona con la separación entre los clusteres. El número de desviaciones estandar por el que el peso de una arista difiera del promedio de los pesos de las aristas cercanas y la proporción de lpeso de la arista con respecto al peso promeido de las aristas cercanas son dos formas de determinar la incosistencia de una arista. Una arista con un factor de incosistencia de dos usualmente enlaza dos clusteres y puede ser borrada. Otras dos estructuras geométricas, el grafo de vecindad relativa y el grafo gabriel, también ha sido usado en el cluster analisis.

grafo de vecindad relativa (RNG) Los patrones x_i y x_j son vecinos relativos, y están conectados en el RNG, si y solo si.

$$d(x_i, x_j) \leq \max\{d(x_i, x_k), d(x_j, x_k)\} \forall k, k \neq i, yk \neq j$$

Podemos decir que x_i y x_j están conectados en RNG sí y solo sí ningún otro punto yace en al intersección de los dos discos de radio $d(x_i, x_j)$ centrados en x_i y x_j.

Grafo Gabriel(GG) Están conectados x_i y x_j sí y solo sí

$$d^2(x_i, x_j) < d^2(x_i, x_k) + d^2(x_j, x_k) \forall k, k \neq i, k \neq j$$

dos puntos están conectados en GG si y solo sí, ningún otro punto yace en el disco de diámetro $d(x_i, x_j)$.

Triangualción de Dlaunay (DT) La definición de la triangulación de Delaunay se entiene mejor en términos de su estructura dual, la teselación de Dirichlet. La teselación de Dirichlet, también conocido como el diagrama de Voronoi, de un conjunto de patrones $H \in R^d$ es una partición de R^d en celdas alrededor de cada patrón x_i tal que cada celda consiste en aquellos puntos de R^d que yacen mas cerca de x_i que de cualquiera de los otros patrones en H. Los patrones de las celdas son intersecciones de las bisectrices perpendiculares de las linesa que conectan a x_i a cada uno de los $n-1$ patrones en H. Por lo tanto cada celda es un polígono convexo. La triangulación de Delauny se define del siguiente modo: la aristas que conecta a los puntos x_i y x_j están en DT sí y solo sí las dos celdas de la teselación de Dirchlet que contiene x_i y x_j comparten una frontera común.

Un arbol recubridor mínimo y la DT juegan un rol importante como fronteras en el RNG y el GG. Puede ser demostrado que

$$E(MST) \subseteq E(RNG) \subseteq E(GG) \subseteq E(DT)$$

donde E denota el conjunto de aristas del grafo. El primera inclusión garantiza que el RNG es un supergrafo del MST. Por lo que todo RNG es conexo

Clustering por vecinos más próximos

Un modo natural de definir clusters es utiilizando la propiedad de vecinos más próximos: un patrón usualmente será puesto en en mismo cluster que su vecino mas cercano. Dos patrones deben ser considerados similares si comparten vecinos. Un algoritmo muy simple que está basado en la regla de vecino más próximo fue propuesto por Lu y Fu [**?**]. Un conjunto de patrones $H = \{x_1, x_2, ..., x - n\}$ tiene que ser particionado en K clusters. El usuario especifica un umbral, t, para la distance entre vecinos.

1. i=1.k=1. Asignar x_1 al custer c_1.

2. i=i+1. Hallar el vecino más próximo de x_i entre los patrones ya asignados a clusters. Sea d_m la distancia de x_i al vecino más próximo. Supongamos que el vecino más próximo está en el cluster m.

3. Si $d_m \leq t$, asignar x_i a C_m. Caso contrario $k = k + 1$ y x_i es un nuevo cluster c_k.

4. Si cada patrón fue asignado a un cluster parar. Sino ir (2).

El número de clusters generados, K, es una función del parámetro t. A medida que el valor de t aumenta, se generan menos clusters. La distancia del vecino más próximo del paso 2 puede ser reemplazado por la distancia promedio entre x_i y su p vecino más proximo en el cluster m. Entonces el usuario tiene que especificar otro parámetro, p. [**?**] utilizó este algoritmo de clustering.

En [**?**] definieron una medida de proximidad como el número de coincidencias en las listas de vecinos más próximos para dos patrones. Su algoritmo de clustering puede ser resumido de la siguiente manera: Ubicar a x_i y x_j en el mismo cluster si x_i y x_j comparten al menos k_t vecinos más próximos y x_i y x_j son k-vecinos más próximos enetre sí.

La noción de proximidad basada en los vecinos más próximos compartidos ha sido modificada por [**?**] mara medir la cercanía mutua de dos patrones. Si x_j es el p-ésimo vecinos mas próximo de x_i y x_i es el q-ésimo vecino más próximo de x_j, entonces el valor de vecindad relativa (MNV) entre x_i y x_j se define como $(p + q)$. Un valor pequeño indica que los patrones son más similares.

Algoritmo de clustering de vecindad mutua

Paso 1 Determinar los k vecinos más próximos de cada patrón.

Paso 2 Calcular el MNV para cada par de patrones. Si dos patrones x_i y x_j no son vecinos mutuos para un valor de k dado, poner $MNV(x_i, x_j)$ a un valor alto.

Paso 3 Identificar todos los pares de patrones con $MNV = 2$. Unir cada uno de esos en un cluster, comenzando con aquellos pares que tengan menor distancia. Repetir el paso 3 para los umbrales de $MNV = 3, 4, ..., 2k$

El parámetro k que controla la profundidad de la vecindad es crucial para la performance del algoritmo. Pequeños valores de k dan muchos clusteres "fuertes" y grandes valores de k dan menos clusteres "debiles". K puede ser elegido lo suficientemente grande para que el algoritmo retorne un solo clusters. En [**?**] demostraron que el algoritmo puede identificar clusteres no esféricos, clusteres linealmente no separables, clusteres con poblacion desigual, y clústeres con puentes de baja densidad cuando $k = 5$ en dos dimensiones.

2.0.17. Metodología

Análisis de datos exploratorios

Vemos el proceso como un loop sin fin en el que nuevos conocimientos y nuevas ideas generan en cada iteración del loop. El resultado final puede ser el diseño de un experimiento que usa herramientas estadísticas estandares para tomar decisiones acerca del fenómeno estudiado. Uno puede derivdad suficiente información acerca del fenómeno a partir de un análisis exploratorio de los datos para extraer conclusiones preliminares.

1. Colección de datos: La obtención cuidadosa de datos en concordancia con lo estandares en el area de aplicación es el primer paso importante en el análisis. La cantidad y los tipos de datos obtenidos influenciará las estrategias disponibles para analizar los datos.

2. Initial screning. Los datos crudos usalmente necesitan cierto tipo de tratamiento antes de que estén listos para el análisis formal.

3. Representación. El problema aqui es poner a los datos en forma adecuada para el analisis posterior. Esto incluye la selección de un indice de proximidad, proyectar los datos a un espacio de características adecuado, examinar la dimensionalidad intrínseca y hacer escalados multidimensionales. El resultado de esta etapa debería ser una matriz de patrones o una matriz de proximidad. La repreesntación elegida dependerá de los datos, el area de aplicación, la experiencia del investigador.

4. Tendencia de clustering. Los datos son aleatorios o existe alguna justificación para el clustering? Esta etapa es, por lo general, ignorada. La información ganada en esta etapa puede, no solo prevenir la aplicación inapropiada de los algoritmos de clustering, sino también proveer información sobre la naturaleza fundamental de los datos.

5. Estrategia de clustering. Una pregunta importante es la elección entre los procedimientos jerárquicos y particionales.

6. Validación. La validación prudente de los resultados del clustering es un paso escencial que transforma el análisis cuantitativo en evidencia. Los indices extternos de validez de clusters comparan los resultados del analisis de clustering con respecto a lo que el investigador desería ver. Los índices internos aseguran el mérito de los resultados de clustering desde una base objetiva. La validación muchas veces incluye análisis de monte carlo y testeo estadístico.

7. Interpretación.

Capítulo 3

Validez de clusters

3.0.18. Introducción

La validez de clusters refiere al procedimiento de evauluar los resultados del análisis de clusters en una forma objetiva y cualitativa. Esto se basa en la premisaque el problema de validez de clusters es inherentemente estadístico. Una estructura de clustering es válida si es inusual en algún sentido. Elegimos expresar el caracter inusual en un framework estadístico y quererir que las probabilidades tengan una interpretación objetiva.

3.0.19. Test de hipótesis

Es sencillo proponer índices de validez de clusters. Es complicado fijar umbrales en esos índices que determinen cuan grande o pequeño debe ser para ser inusual. Los métodos estadísticos proveen un framework para decidir racionalmente cuan grande es "grande" y cuan chico es "chico".

Una estadística T es una función de los datos que, se supone, contiene información útil. Un estadístico T puede ser el error cuadrático de un agrupamiento o el nivel en el que la partición forma una jerarquía o una medida de compacidad para un cluster. En términos matemáticos, T es una variable aleatoria y su distribución descbire la frecuencia relativa en el que los valores de T ocurren bajo alguna hipótesis. Una distribución requiere que exista un espacio muestral. Una hipótesis es una sentencia acerca de la frecuencia relativa de los eventos en el espacio muestral que expresan el concepto de frases como "los datos son aleatorios" o "los datos están clusterizados". Una hipótesis es testeada observando los valores de T y decidiendo si la obvservación es inusual, basada en al distribución de T.

Hipótesis aleatoria

La hipótesis nula en validez de clusters es una sentencia sobre "no estructura", o aleatoriedad, que afirma la frecuencia con que se producen los

miembros de una población de referencia. Las tres hipótesis nulas más comunes en en trabajo de validez de clusters son las hipótesis de grafos aleatorios, la hipótesis de etiquetas aleatorias y la hipótesis de posiciones aleatorias. El subíndice "0" refiere a la hipótesis nula. Un gran problema en la validez de clusters es establecer las distribuciones de los estadísticos bajo la hipótesis nula.

Hipótesis de grafos aleatorios Es utilizada usualmente cuando solo se encuentra disponible información interna, es decir, información que concierne solo a los vectores o sus relaciones. Es apropiado cuando se utilizan proximidades ordinales entre los vectores. Antes de proceder, definamos la matriz ordinal, la matriz orden de rango, $X \times N$ como una matriz simétrica con los elementos de la diagonal en 0, (siempre que se utilicen medidas de desemejanza) y con los elementos de la matriz diagonal superior siendo enteros en el rango $[1, N(N-1)/2]$. El elemento $A(i,j)$ de A provee solo información cualitativa acerca de la desemekanza entre los vectores x_i y x_j. Si, por ejemplo $A(2,3) = 3$ y $A(2,5) = 5$, solo podemos concluir que x_2 es mas similar a x_3 que a x_5. Sea A_i una matriz de orden de rango $N \times N$ sin valores iguales. Bajo la hipótesis del grafo aleatorio, la población de referencia consiste en aquellas matrices A_i cada una generada insertando aleatoriamente los enteros en el rango $[1, \frac{N(N-1)}{2}]$, es sus elementos de la matriz diagonal superior. Sea P la matriz de proximidad ordinal asociada a un conjunto de datos X y C una estructura de agrupamientos producida pro la aplicación por medio de un algoritmo específico P. Finalmente, sea C_i la estructura de agrupamientos producida cuando el mismo algoritmo es aplicado a A_i. Definimos un estadístico q que mida el acuerdo entre la matriz de orden de rango y la estructura del cluster. Si el valor de q, correspondiente a Py C,es inusualmente grande o pequeña, la hipótesis nula es descartada.

Hipótesis de etiquetas aleatorias Consideremos todas las posibles particiones, P', de X en m grupos. Cada partición puede ser definida en términos de un mapeo g desde X hasta $\{1, ...m\}$. La hipótesis de etiquetas aleatorias asume que *todos* los posibles mapeos son igualmente probables. El estadístico q puede ser definido a fin de medir el grado en que la información inherente en el conjunto de datos X, tales como la matriz de proximidad P, coincide con una partición específica. El estadístico q se utiliza entonces para probar el grado de coincidencia entre P y una partición P impuesta externamente, en contra de la q_i particiones que corresponden a las particiones aleatorias generados bajo la hipótesis de etiquetas aleatorias. Una vez más, se rechaza H_0 entonces si q es inusualmente grande pequeña.

Hipótesis de posiciones aleatorias Esta hipótesis es apropiada para da-

tos proporcionales. Requiere que "todos las configuraciones posibles de los N vectores en una región específica del espacio l-dimensional puedan ocurrir con la misma probabilidad". Estas regiones pueden ser le hipercubo H_l o la hiperesfera l-dimensional. Una forma de producir una de estas configuraciones es insertar cada punto aleatoriamente en estas regiones, de acuerdo a una distribución uniforme.

3.0.20. Definición de un test

Supongamos que se ha acordado un estadístico T y una hipótesis nula H_0. Supongamos que la distribución de T es conocida bajo la hipótesis nula (encontrar esta distribución es muy complicado). ¿Como probar si la hipótesis nula es una descripción apropiada para los datos? En la validez de clusters queremos determinar si una jerarquía o una partición de los datos es inusualmente buena. Para seer inusual, el ajuste de estos agruapmientos debe ser al menos mejor que el ajuste de la jerarquía o el agrupamiento a un conjunto de datos aleatorio.

Sea $P(B|H_0)$. El evento B podría ser "$T \geq t$" o "$T \leq t$", donde t es un umbral. Uno debe cuando un valor grande o pequeño de T corresponde a una buena estructura. Por ejemplo, un error cuadrático pequeño, indica un mejor agrupamiento, pero la proporción entre clusters con respecto a la intracluster debe ser grande para corresponderse con un buen agruapamiento.

Sea α sea un número pequeño, como 0,05 o 0,01, llamado tamaño o nivel de un test. Dada la distribución de T bajo H_0, y asumiendo que un valor grande de T indica que H_0 debe ser rechazada, podemos ubicar un umbral, t_α, en T resolviendo la siguiente ecuación.

$$P(T \geq t_\alpha | H_0) = \alpha$$

Supongamos que el valor de T medido en un experimento es t^*. Para responder la pregunta "¿Debe ser rechazada H_0?" puede ser replanteada como "¿Es t^* lo suficientemente grande para llamarlo inusual?"

$$\text{Si } t^* \geq t_\alpha, \text{ rechazar } H_0 \text{ al nivel } \alpha$$

Entonces α es la probabilidad de decididir en contra de H_0 cuando H_0 es verdadera. Este test produce una de dos respuestas, o bien T es grande o no lo es, de forma que la estructura del agruapemiento es válida a nivel α o no lo es. La región crítica del test es el conjunto de valores del estadístico que conducen al rechazo de H_0, o $\{t : t \geq t_\alpha\}$.

3.0.21. Potencia de un test

El test de H_0 cuenta solo la mitad de la historia. Lo que está faltando es la hipótesis de una estructura, o hipótesis alternativa, H_1, con la cual comparar con H_0. Las hipótesis alternativas establecen estructuras específicas, tales

como “los clusters son muestras de una mixtura Gaussiana” o “ los datos contienen dos clusters”. Si la región crítica de un test es $\{t : t \geq t_\alpha\}$, el poder del test es la probabilidad de alcanzar la decisión correcta cuando H_1 es verdadera, o

$$poder = P(T \geq t_\alpha | H_1)$$

Buscamos estadísticos T que conduzcan a tests con gran potencia.

En conclusión, nuestro enfoque para la validez de clusters involucra varios pasos. Hay que definir una hipótesis nula que expresa nuestra idea de no estructura teniendo en cuenta el tipo de dato. Tiene que ser seleccionado un estadístico, o índice, sensible a la presencia de estructuras en los datos y se debe establecer la distribución del estadístico bajo la hipótesis nula. Un umbral puede ser hallado que define cuand grande es “grande” para el estadístico seleccionado. El umbral establece un test de hipótesis formal. La potencia del test evalua la habilidad del estadístico en reconocer la presencia de la estructura especificada en la hipótesis alternativa.

3.1. Estadístico Γ de Hubert

Una forma de validar una estructura de clustering es compararla con una estructura a priori, que es asignada sin importar las mediciones. El estadístico que se conoce como Γ de Hubert (Γ_H)es efecitvo en la evaluación del ajuste entre datos y estructuras a priori [**?**]. El problema abstracto en el que Γ_H es aplicable puede ser definido como. Sea $H = [X(i,j)]$ e $Y = [Y(i,j)]$ dos matrices de proximidad $n \times n$ en los mismos n objetos. Por ejemplo $X(i,j)$ puede denotar la proximidad observada entre los objetos i y j e $Y(i,j)$ puede ser definida como:

$$Y(i,j) = \begin{cases} 0 & \text{Si los objetos } i \text{ y } j \text{ tienen la misma etiqueta} \\ 1 & \text{Si no} \end{cases}$$

Otras aplicaciones tienen $X(i,j)$ como la separación geográfica entre los objetos i y j, mientras que $Y(i,j)$ es la separación de esos objetos en tiempo. El estadístico Γ de Hubert es, simplemente, la correlación serial de puntos (point serial correlation) entre las dos matrices. Si las dos matrices son simétricas pueden ser expresado como:

$$\Gamma = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} X(i,j)Y(i.j)$$

En la forma normalizada, Γ son los coeficientes de correlación de muestras entre los elementos de las dos matrices. Si m_x y m_y denotan las medias muestrales y s_x y s_y denotan la desviaciones estandar muestrales de los elementos de las matrices H e Y, el estadístico Γ normalizado es:

$$\Gamma = \frac{\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}[X(i,j)-m_x][Y(i,j)-m_y]}{s_x s_y}$$

donde $M = \frac{n(n-1)}{2}$ es el numero de entradas en la doble suma y los momento están dados por

$$\begin{array}{cc} m_x = \frac{1}{M}\sum\sum X(i,j) & m_y = \frac{1}{M}\sum\sum Y(i,j) \\ s_x^2 = \frac{1}{M}\sum\sum X^2(i,j) - m_x^2 & s_y^2 = \frac{1}{M}\sum\sum Y^2(i,j) - m_y^2 \end{array}$$

Todas las sumas son sobre el conjunto $\{(i,j) : i \leq i \leq n-1, i+1 \leq j \leq n\}$. El estadístico Γ mide el grado de correspondencia lineal entre las entradas de H e Y. Valores inusualmente grandes de Γ sugieren que las dos matrices coindicen etre sí. El estadístico Γ normalizado está siempre en -1 y 1. La aplicación más comun de Γ comprueba si la asociación entre H e Y es inusualmente grande bajo la hipótesis de las etiqueatas aleatorias. Es decir, pueden que las filas y columnas de unas de las matrices hayan sido insertadas de forma aleatoria. La hipótesis de etiquetas aleatorias puede ser formalizado de la siguiente manera: H_0 Todas las permutaciones de las etiquetas de las filas (o columnas) de $[Y(i,j)]$ son igualmente probables. La permutación referenciada en H_0 es un reordenamiento de las etiquetas de los objetos $\{1,2,....n\}$. La reordenación se aplica a a las filas y columnas de Y, de formas que Y queda mezclada.

3.1.1. Ejemplo

$$H = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{pmatrix} - & 1,2 & 0,6 & 0,2 \\ - & 0 & 0,3 & 0,4 \\ - & - & 0 & 0,1 \\ - & - & - & 0 \end{pmatrix} \end{array} \quad Y = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \end{array} \begin{array}{c} \begin{array}{cccc} x_1 & x_2 & x_3 & x_4 \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 1 \\ - & 0 & 1 & 0 \\ - & - & 0 & 1 \\ - & - & - & 0 \end{pmatrix} \end{array}$$

La matriz H es una matriz de desemejanza para cuatro objetos y la matriz Y se deriva ubicando a los objetos x_1 y x_3 en una categoría y los objetos x_2 y x_4 en una segunda categoría, luego asignando una proximidad 0 a objetos en la misma categoría y proximidad 1 para objetos en diferentes categorías. Hay 24 permutaciones de la secuencia de números$\{1,2,3,4\}$ en los objetos. La distribución de Γ bajo H_0 se obtiene calculando:

$$\sum_{i=1}^{3}\sum_{j=i+1}^{4} X(i,j)Y[g(i),g(j)]$$

para cada una de las 24 permutaciones $\{g(1),g(2),g(3),g(4)\}de\{1,2,3,4\}$.Solo los tres valores moestrados a contnuación puden ocurrir.

Valor de Γ	1,5	1,8	2,3
Frecuencia	8	8	8

El valor observado de Γ, calculao a partir de H y la matriz Y original, es 1.8. Entonces un valor como el obserado no es inusual y podemos concluir que las filsa y las columnas de Y ffueron etiquetadas de forma aleatoria. La realidad computacional nos impide aplicar este procedimiento para problemas prácticos. Con 8 objetos tenemos $8! = 40320$ valores de Γ. Es claro que necesitamos tenemos que aproximar o estimar la distribución de Γ bajo H_0. Los dos enfoques principales para este problema es el análisis de Monte Carlo y la estimación de momentos. Sea $\{g(1), g(2), ..., g(n)\}$ denota la permutación de los enteros $\{1, 2, ..., n\}$. La variable aleatoria cuya distribución es requerida bajo H_0 puede ser escrita como:

$$\Gamma(g) = \frac{\frac{1}{M}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}[X(i,j)-m_x][Y(g(i),g(j)]-m_y]}{s_x s_y}$$

Un análisis de Monte Carlo elige aleatoriamente un numero de permutaciones g, y ordena la secuencia resulante $\{\Gamma(g)\}$ para estimar la distribución de Γ bajo H_0.

3.1.2. Ejemplo

Este ejemplo demuestra la estimación de la distribución de Γ bajo la hipótesis de etiquetas aleatorias para dos matrices de patrones. El dataset 80X, contiene 15 patrones en un espacio 8-dimensional par acada una de las 3 categorías para formar un total de 45 patrones. Supongamos que queremos comprobar si las etiquetas de categorías coinciden inusualmente bien a las ubicaciones de los patrones en el espacio de características 8-dimensional. La matriz H es la matriz de las distancias euclieas entre los patrones e Y es una matriz que contiene un 1 en la posición (i, j) si el patrón i y j estpan en diferentes categorías y 0 si están en la misma categoría. Si los primeros 15 patrones son de la categoría 8, los otros 15 son de la categoría O, y los últimos 15 están en la categoría X, entonces Y pueden ser bloques de 15×15 submatrices:

$$Y = \begin{bmatrix} 0 & I & I \\ I & 0 & I \\ I & I & 0 \end{bmatrix}$$

$$m_y = \frac{3\times 15^5}{990} = \frac{675}{990} = 0{,}682 \quad \text{y} \quad s_y = [\frac{675}{990} - (\frac{675}{990})^2]^{\frac{1}{2}} = 0{,}4666$$

La media y la desviación estandar de las distancias euclideas entre patrones son:

$$m_x = 9{,}07 \quad \text{y} \quad s_x = 1{,}88$$

El estadístico gamma puede ser escrito de la siguiente forma:

$$\Gamma(g) = \frac{\frac{1}{990}\sum_{i=1}^{44}\sum_{j=i+1}^{45}X[i,j]Y[g(i),g(j)]-(0{,}682)(9{,}07)}{(0{,}466)(1{,}88)}$$