PROGRAMA EDUCATIVO
Maestría en ciencias de la computación

UNIDAD DE APRENDIZAJE
MINERIA DE DATOS

Procesos de análisis no supervisado

ELABORACION
ADRIAN TRUEBA ESPINOSA

PRESENTACIÓN DEL CURSO

La unidad de aprendizaje “Minería de Datos”, se imparte en la Maestría en Ciencias de la Computación. Tiene la finalidad de desarrollar las competencias en los alumnos para aplicar las metodologías para la predicción de datos a partir de algoritmos supervisados y no supervisados

Objetivo del curso

En este curso el estudiante conocerá y aplicara las metodologías para la predicción de datos que permitan pronosticar salidas de datos y revelar sus relaciones a partir de algoritmos empleados en la minería de datos supervisados y no supervisados

Justificación académica

Con el uso de las tecnologías de la información el almacenaje de datos se ha incrementado en las ultimas dos décadas de manera sustancial, estos datos pueden contener conocimiento que a simple vista no se pueden visualizar y entender. Para atacar esta situación se utiliza la minería de datos y con ello poder el conocimiento que pueda ser explotado y ser usado por las empresas o entidades publicas a fin de mejorar procesos o incrementar la entrada de recursos económicos

CONTENIDO DEL CURSO

Unidad I. Métodos para el tratamiento y análisis de datos
Unidad II. Procesos de análisis supervisados
Unida III. Procesos de análisis no supervisados
Unidad IV. Métodos estimadores de error
Unidad VI. Métodos para análisis del índice de aciertos

METAS A ALCANZAR

Que el alumno desarrolle las competencias técnicas y de especialidad para el tratamiento de datos con metodologías de análisis no supervisado con técnicas de agrupamiento

OBJETIVO DEL MATERIAL DIDÁCTICO

Que el alumno conozca el método para predicción y análisis de datos con algoritmos no supervisados (claustering)

METODOLOGÍA DEL CURSO

El curso se desarrollará bajo el siguiente proceso de estudio:

1. Exposición de parte del profesor mediante la utilización de este material en diapositivas.
2. Control de lecturas selectas que el profesor asignará para complementar la clase.
3. Trabajos donde se investigarán temas, conceptos, procesos y métodos de los temas por ver.
4. Participación en clases
5. Desarrollo de trabajo individual en casos de estudio

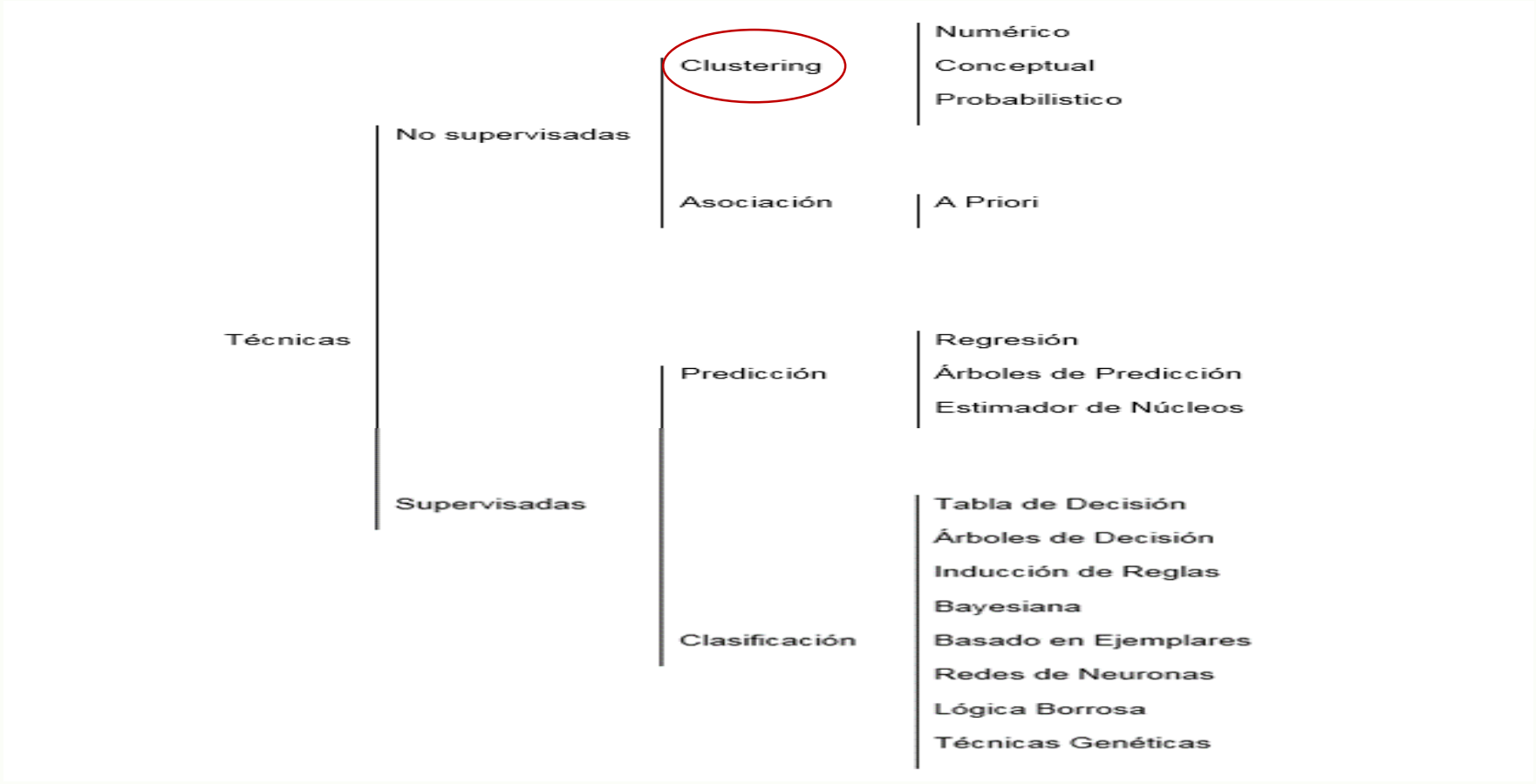
UTILIZACIÓN DEL MATERIAL DE DIAPOSITIVAS

El material didáctico visual es una herramienta de estudio que sirve como guía para que el alumno repase los temas más significativos de “el análisis supervisado”, cabe aclarar que será un tutor el cual proporcionará las ideas generales del tema, asiendo ejercicios en el salón de clase.

UNIDAD DE COMPETENCIA

Proceso de análisis no supervisado

Las técnicas de Minería de Datos se clasifican en dos grandes categorías: supervisadas o predictivas y no supervisadas o descriptivas



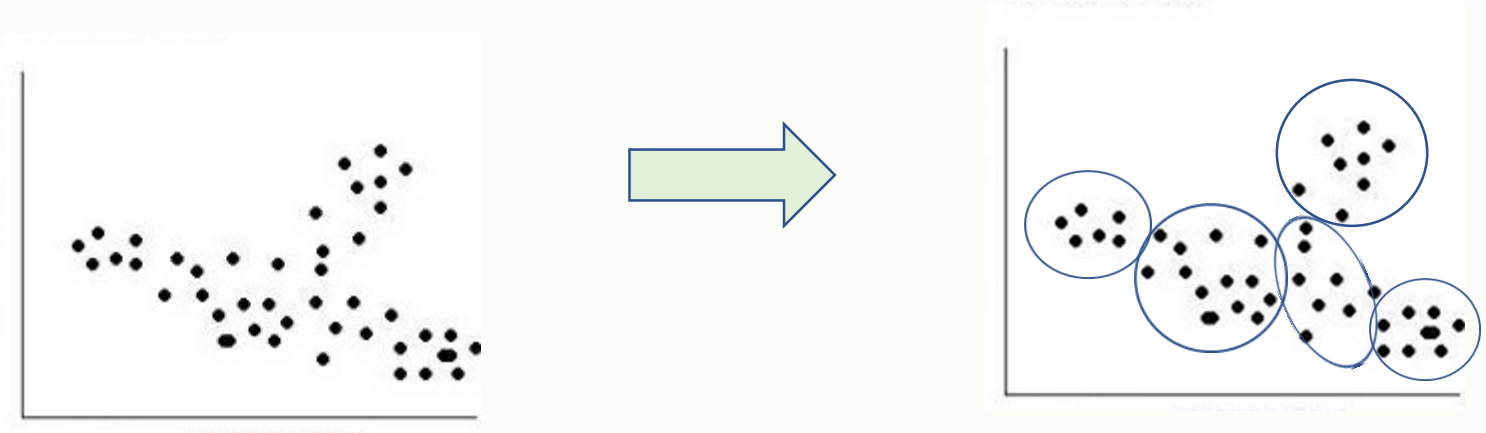
En esta unidad de aprendizaje se abordara las técnicas no supervisadas de clustering con énfasis a los jerárquicos

Análisis de Clustering

También es conocido como Análisis de Conglomerados

El análisis de clustering es una técnica estadística multivariante de clasificación automática de datos. Que busca agrupar un conjunto de observaciones en un número dado de clusters o grupos, tratando de lograr la máxima homogeneidad en cada grupo y la mayor diferencia entre los grupos. Este agrupamiento se basa en la idea de distancia o similitud entre las observaciones.

Los individuos que puedan ser considerados similares sean asignados a un mismo cluster, mientras que individuos diferentes (disimilares) se localicen en clusters distintos.



El análisis cluster define grupos tan distintos como sea posible en función de los propios datos.

Objetivo del agrupamiento

- Dividir un conjunto de n vectores $\mathbf{x}_j$, en c grupos G_i y encontrar el centro de cada grupo.

- Ordenar los vectores en los grupos de acuerdo a similitud de características.

- Minimizar la función de disimilitud.

Estudia tres tipos de problemas:

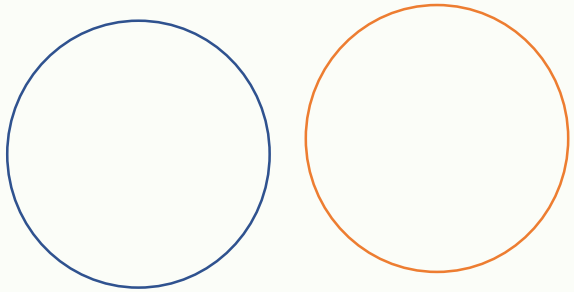
1. Partición de los datos: disponemos de observaciones que pensamos son heterogéneas y deseamos dividirlos en un nº prefijado de grupos, de tal manera que todo elemento quede clasificado y pertenezca a un solo grupo y los grupos sean internamente homogéneos.

2. Construcción de jerarquías: deseamos estructurar los elementos de un conjunto de forma jerárquica por su similitud ⇒ ordenar en niveles.

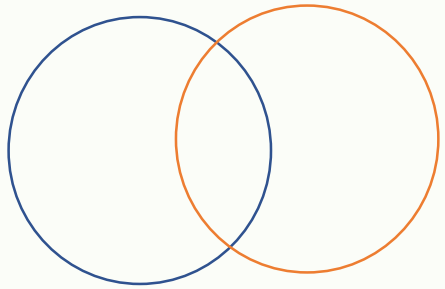
3. Clasificación de variables: en problemas con muchas variables es interesante hacer una división en grupos para luego reducir la dimensión. Longitud cuello (m) Forma manchas Girafas Defini

Principales tipología de agrupamiento

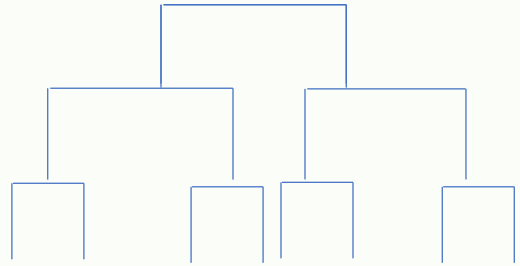
Sin solapamiento



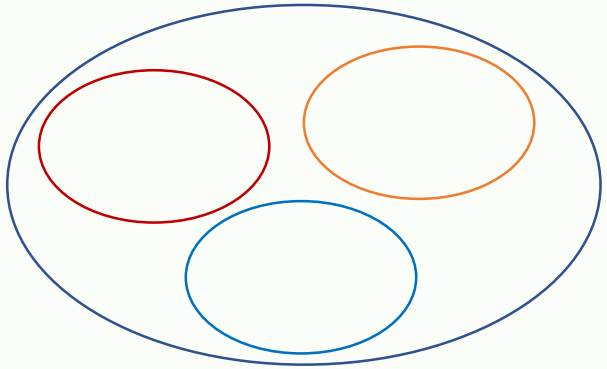
Con solapamiento



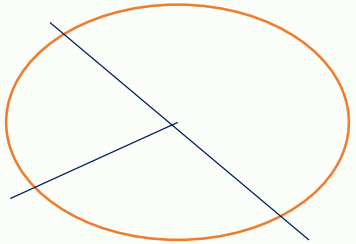
Jerárquico



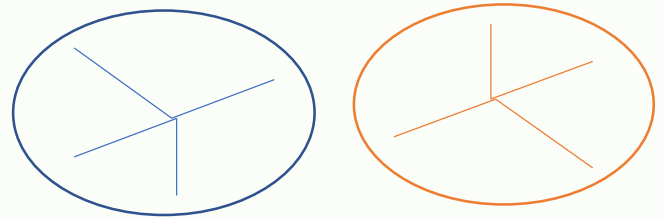
Aglomerativo



Divisivo



No jerárquico



Estimación del modelo

Métodos jerárquicos: Se comienza con tantos grupos como individuos y se van agrupando según diversos criterios:

- centroide
- vecino más cercano
- vecino más lejano...

No jerárquicos:

- Se determina el número de grupos
- Se proporciona un centroide inicial
- Se van incorporando a los individuos hasta que se cumpla un criterio de parada

Selección:

- Ambos
- Primero método jerárquico para establecer número de grupos y centroides iniciales
- Luego método no jerárquico

Diferencias básicas

Métodos jerárquicos	Métodos no jerárquicos:
Comienza con las observaciones y no precisa determinar a priori el número de conglomerados	Comienza con una partición inicial de conglomerados. A priori se determina el número y composición de los conglomerados
La asignación de objetos es definitiva	El procedimiento es iterativo y permite la reasignación de objetos
Operan con una matriz de similaridades	Operan con la matriz de datos originales

Inconvenientes principales

Métodos jerárquicos	Métodos no jerárquicos
Si la estructura de la muestra es desconocida resulta difícil escoger el algoritmo	Dificultad en conocer a priori el número real de los conglomerados existentes en la muestra
Es difícil operar e interpretar los gráficos con más de 200 datos.	Formar todas las particiones posibles para escoger la óptima es muy complejo
Mayor cantidad de atípicos en esta partición	Mayor complejidad en los análisis.
Una mala partición no puede modificarse	Una mala decisión inicial sobre el nº y composición de los grupos ocasiona una errónea clasificación

ETAPAS EN UN ANÁLISIS DE CLUSTER

se necesite identificar segmentos o agrupar los datos en grupos homogéneos, se recomienda seguir el siguiente procedimiento:

1. Formular el problema. En esta etapa, el investigador puede realizar entrevistas informales para identificar y seleccionar las variables en las que basará la agrupación, evitar variables irrelevantes.

2. Seleccionar una medida de similitud. Se necesita una forma de medir la diferencia o semejanza entre observaciones u objetos, la forma en que generalmente se hace es en términos de la distancia entre cada par de casos; cuando la distancia es menor se considera que los casos son más parecidos entre sí.

3. Seleccionar un procedimiento de agrupamiento.

4. Decidir el número de conglomerados a conservar. Una vez que ya se ha hecho un clasificación, se decidirá con cuántos conglomerados se trabajará o en cuántos segmentos se dividirá.

5. Interpretar y elaborar un perfil de los conglomerados. En esta etapa se procederá a determinar las características de cada conglomerado que se conservará

ALGORITMOS DE CONGLOMERACIÓN

De la elección del algoritmo de clasificación dependen el número y composición de los conglomerados obtenidos.

El algoritmo es la forma particular de cálculo empleado.

La elección del algoritmo de clasificación depende de:

a) Los objetivos del estudio
b) Las características de los datos: métrica de las variables y tamaño muestral
c) El método elegido: jerárquico o no jerárquico
d) Los límites del programa y ordenador que usemos

Los algoritmos tienen siempre la misma estructura y sólo se diferencian en la forma de calcular las distancias entre grupos

1. Empezar con tantas clases como elementos, n
2. Seleccionar los dos datos más próximos y formar con ellos una clase
3. Sustituir los dos elementos anteriores por uno sólo que representa a la clase
4. Se calculan las distancias entre éste nuevo elemento y los anteriores
5. Repetir 2 y 3 hasta agrupar todos los datos en una sola clase

Existen diversas maneras de calcular la distancia, las que se aplican con mayor frecuencia son:

La distancia euclidiana que es la raíz cuadrada de la suma de las diferencias al cuadrado entre los valores de dos casos para cada variable;

La distancia de Manhattan o de calles urbanas entre dos casos es la suma de los valores absolutos de la diferencia entre observaciones para cada variable;

La distancia de Chebychev entre dos objetos es el valor absoluto de la diferencia máxima entre los valores para cualquier variable2 .

Existen otras como: La de Mahalanobis, **la de Minkowski**, **la de Tchebychef**

ELECCIÓN DE MEDIDAS DE DISTANCIA Y SIMILARIDAD

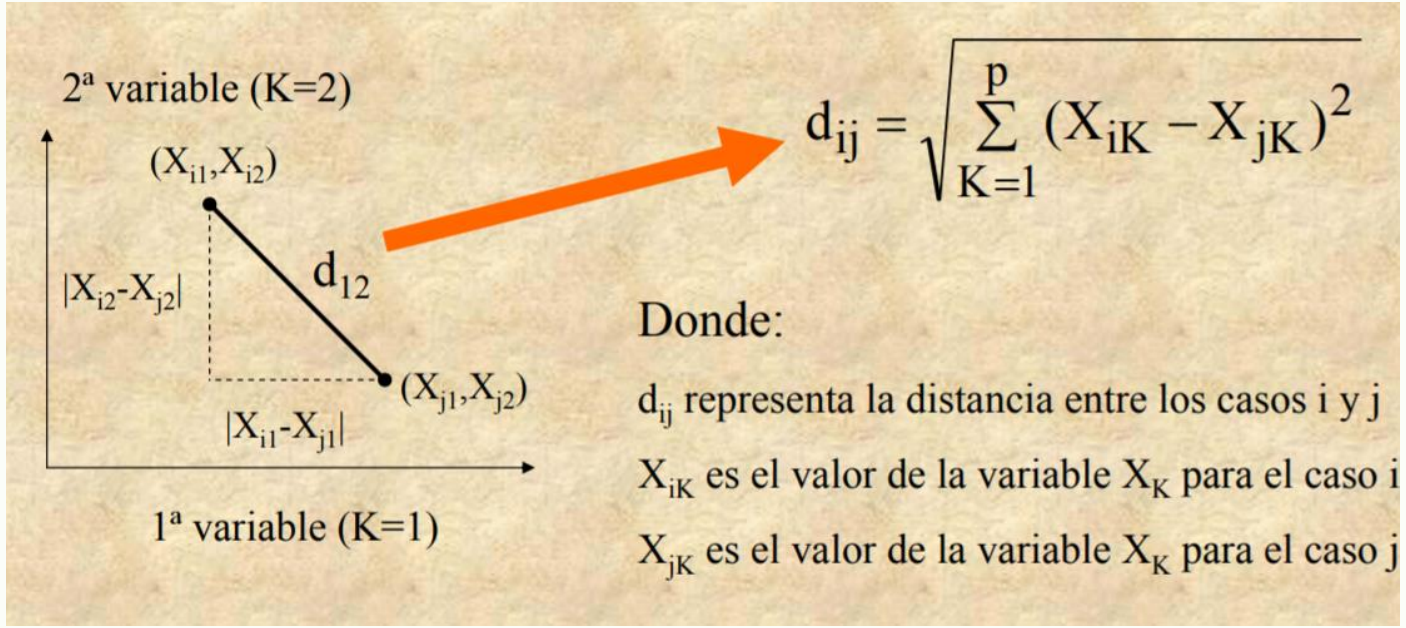
Los criterios para decidir qué objeto se incluye o no en un conglomerado se utilizan matrices de distancias o similaridades entre los pares de objetos.

Las más empleadas para **variables cuantitativas** son las distancias euclídea, euclídea al cuadrado, “city block” y la correlación.

Las más empleadas para **variables binarias** son la distancia euclídea junto con el coeficiente de Jaccard. La más empleada para variables cualitativas es la chi-cuadrado.

Medidas de similaridad para variables métricas

Distancia euclídea (d)



$$d_{ij} = \sqrt{\sum_{K=1}^{p} (X_{iK} - X_{jK})^2}$$

Donde:

d_{ij} representa la distancia entre los casos i y j

X_{iK} es el valor de la variable X_K para el caso i

X_{jK} es el valor de la variable X_K para el caso j

Euclídea al cuadrado

Empleada por defecto para datos de intervalo en especial cuando se agrupan casos

Medida recomendada en el algoritmo del centroide y de Ward En la que más influyen las diferencias en las medidas

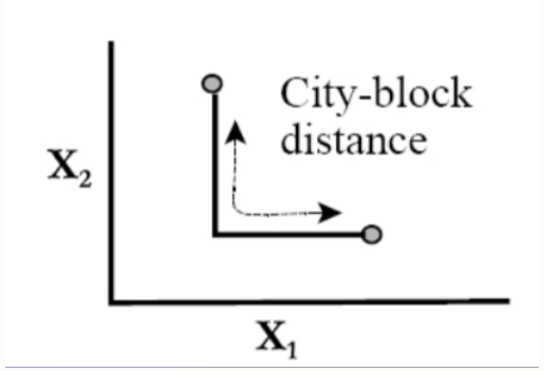
$$d_{ij}^2 = \sum_{K=1}^{p} (X_{iK} - X_{jK})^2$$

Manhattan o “city-block”

Las mayores medidas de disimilitud ecológicas son del tipo Manhattan.

- Comparadas a la euclídea, le dan menos peso a los outliers (no están diferencias al cuadrado)
- Comparada a DE, retienen la sensibilidad al incrementar la heterogeneidad en el conjunto de datos.
- No es para distancias no proporcionales.

$$d_{ij} = \sum_{K=1}^{p} \left| X_{iK} - X_{jK} \right|$$



Correlación

Se aplica a variables continuas, y usa correlaciones (Pearson, Spearman o Kendall). También se emplea en métodos para jerárquizar variables.

$$d_{jh} = \sqrt{\frac{1 - r_{jh}}{2}}$$

Coeficiente de Jaccard

Conocido como razón de similaridad, se aplica a variables binarias.

Objeto i \ Objeto j	1	0
1	a	b
0	c	d

$$s_{ij} = \frac{a}{a+b+c}; \qquad d_{ij} = \sqrt{2(1-s_{ij})}$$

Chi-cuadrado

$$d_{jh} = \sqrt{\chi^2}$$

usa coeficiente de contingencia para variables binarias

Distancia Minskowski

Generalización de la distancia Euclidiana mediante el parámetro r

$$dist = \left(\sum_{k=1}^{n} | p_k - q_k |^r \right)^{\frac{1}{r}}$$

r = 1. Distancia Manhattan
Ejemplo típico: Distancia de Hamming: Numero de bits diferentes entre dos arreglos de bits

r = 2. Distancia Euclidiana r → ∞. Distancia “supremo” (norma Lmax o L∞). La máxima diferencia entre los atributos

Aplicaciones de agrupamiento

- También conocido como C-means, ha sido aplicado en áreas como:
  - Procesamiento de imágenes.
  - Compresión de datos.
  - Procesamiento de voz.
  - Entre otros.

Métodos no jerárquicos

K-means

- Usualmente se emplea distancia Euclidiana para medir la disimilitud. La función de disimilitud se define como:

$$J = \sum_{i=1}^{c} \sum_{\mathbf{x}_k \in G_i} \|\mathbf{x}_k - \mathbf{c}_i\|^2$$

- Los grupos particionados se definen por una matriz de $c \times n$ de pertenencia $\mathbf{U}$, donde el elemento $u_{i,j}$:
  - Es 1 si el vector $\mathbf{x}_j$ pertenece al grupo i,
  - Es 0 en caso contrario.

K-means

- En otras palabras:

$$u_{i,j} = \begin{cases} 1, & \left\|\mathbf{x}_j - \mathbf{c}_i\right\|^2 \leq \left\|\mathbf{x}_j - \mathbf{c}_k\right\|^2, \forall k \neq i \\ 0, & \text{en caso contrario} \end{cases}$$

- Ya que un punto puede pertenecer solamente en un grupo, la matriz de pertenencia $\mathbf{U}$ tiene la siguiente propiedad:

$$\sum_{i=1}^{c} u_{i,j} = 1, \forall j = 1, \ldots, n$$

K-means

- Tambien se cumple que:

$$\sum_{i=1}^{c}\sum_{j=1}^{n} u_{i,j} = n$$

- Si $u_{i,j}$ es fijo, el centro optimo $\mathbf{c}_i$ que minimiza J es la media de los vectores del grupo i:

$$\mathbf{c}_i = \frac{1}{|G_i|}\sum_{\mathbf{x}_k \in G_i} \mathbf{x}_k \qquad \textbf{Ec. (1)}$$

K-means

- Donde $|G_i| = \sum_{j=1}^{n} u_{i,j}$ es el tamaño de G_i.

- El algoritmo de K-means es presentado con un conjunto de datos $\mathbf{x}_i$, $i = 1, \ldots, n$.

- El algoritmo determina los centros de los grupos $\mathbf{c}_i$, y de la matriz de pertenencia $\mathbf{U}$ iterativamente empleando los siguientes pasos:

K-means algoritmo

1. Inicializar los centros de los grupos $\mathbf{c}_i$, $i = 1, \ldots, c$. Típicamente se seleccionan aleatoriamente c puntos del conjunto de datos.

2. Determinar la matriz de membresía $\mathbf{U}$ con la función de pertenencia.

3. Calcular la función de costo de acuerdo a la función J. Detener si se alcanza un valor de tolerancia o si su desempeño sobre iteraciones previas es menor a un cierto umbral.

4. Actualizar los centros de los grupos empleando la ecuación 1. Ir al paso 2

Ejemplo

- Sean los vectores:

$$x_1 = \begin{bmatrix} 10 \\ 0 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 10 \\ 10 \\ 0 \end{bmatrix}, x_3 = \begin{bmatrix} 9 \\ 7 \\ 1 \end{bmatrix}, x_4 = \begin{bmatrix} 3 \\ 8 \\ 5 \end{bmatrix}$$

- Se inicializa aleatoriamente la matriz de pertenencia y los centros de los grupos, en este ejemplo se busca crear dos grupos:

$$U = \begin{bmatrix} 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{bmatrix}, c_1 = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix}, c_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Ejemplo

- Se calcula cada elemento de la matriz de pertenencia:

$\|x_1 - c_1\|^2 \leq \|x_1 - c_2\|^2$, entonces $u_{1,1} = 1$

$\|x_2 - c_1\|^2 \leq \|x_2 - c_2\|^2$, entonces $u_{1,2} = 1$

$\|x_3 - c_1\|^2 \leq \|x_3 - c_2\|^2$, entonces $u_{1,3} = 1$

$\|x_4 - c_1\|^2 > \|x_4 - c_2\|^2$, entonces $u_{1,4} = 0$

- Así sucesivamente, en la primera iteración la matriz queda como:

$$U = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

- Se calculan los centros de los grupos, pero nótese que $G_1 = \{x_1, x_2, x_3\}$ y que $G_2 = \{x_4\}$

Ejemplo

- Entonces $|G_1| = 3$ y $|G_2| = 1$, por lo tanto:

$$c_1 = \frac{1}{3}\begin{bmatrix}29\\17\\1\end{bmatrix}, c_2 = \begin{bmatrix}3\\8\\5\end{bmatrix}$$

- Se calcula la función de similitud:

$$J = 32.34 + 19 + 2.67 + 0 = 54.01$$

- Se vuelven a realizar todos los cálculos hasta minimizar la función de similitud.

Fuzzy C-means

- Es un algoritmo de agrupamiento de datos en el que cada punto tiene un grado de pertenencia en cada grupo.

- El algoritmo divide una colección de n vectores $\mathbf{x}_j$, $j = 1, \ldots, n$ en c grupos difusos y se encuentran los centros de cada grupo, tal que la función de disimilitud es minimizada.

Fuzzy C-means

- La función de disimilitud se define como:

$$J = \sum_{i=1}^{c} \sum_{j=1}^{n} u_{i,j}^{m} \left\| \mathbf{c}_i - \mathbf{x}_j \right\|^2$$

- Donde $u_{i,j} \in [0,1]$ es el grado de pertenencia del vector $\mathbf{x}_j$ en el grupo i; $\mathbf{c}_i$ es el centro del grupo i; y $m \in [1,\infty)$ es la ponderación del exponente.

Fuzzy C-means

- La matriz de pertenencia $\mathbf{U}$ debe cumplir que:

$$\sum_{i=1}^{c} u_{i,j} = 1$$

- Para todo $j = 1, \ldots, n$

Fuzzy C-means

- Calculo de centros:

$$\mathbf{c}_i = \frac{\sum_{j=1}^{n} u_{i,j}^{m} \mathbf{x}_j}{\sum_{j=1}^{n} u_{i,j}^{m}}$$

- Matriz de pertenencia:

$$u_{i,j} = \frac{\left\|\mathbf{x}_j - \mathbf{c}_i\right\|^{-2/(m-1)}}{\sum_{k=1}^{c} \left\|\mathbf{x}_j - \mathbf{c}_k\right\|^{-2/(m-1)}}$$

Fuzzy C-means

1. Inicializar la matriz $\mathbf{U}$ con valores aleatorios entre 0 y 1.

2. Calcular los c centros de todos los grupos.

3. Calcular la función de disimilitud. Detener si su valor es menor a un valor de tolerancia definido.

4. Calcular nuevamente la matriz de pertenencia $\mathbf{U}$. Ir al paso 2.

Ejemplo

- Sean los vectores:

$$x_1 = \begin{bmatrix} 10 \\ 0 \\ 0 \end{bmatrix}, x_2 = \begin{bmatrix} 10 \\ 10 \\ 0 \end{bmatrix}, x_3 = \begin{bmatrix} 9 \\ 7 \\ 1 \end{bmatrix}, x_4 = \begin{bmatrix} 3 \\ 8 \\ 5 \end{bmatrix}$$

- Se inicializa aleatoriamente la matriz de pertenencia y los centros de los grupos, en este ejemplo se busca crear dos grupos:

$$U = \begin{bmatrix} 0.7 & 0.2 & 0.3 & 0.4 \\ 0.3 & 0.8 & 0.7 & 0.6 \end{bmatrix}, c_1 = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix}, c_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$$

Ejemplo

- Se calcula cada elemento de la matriz de pertenencia:

$$u_{1,1} = \frac{\|x_1 - c_1\|^{-2}}{\sum_{k=1}^{2} \|x_1 - c_k\|^{-2}}$$

Donde $\sum_{k=1}^{2} \|x_1 - c_k\|^{-2} = \|x_1 - c_1\|^{-2} + \|x_1 - c_2\|^{-2} = 0.021$

Entonces:

$$u_{1,1} = \frac{0.02}{0.021} = 0.95$$

Ejemplo

$$u_{2,1} = \frac{\|x_1 - c_2\|^{-2}}{\sum_{k=1}^{2} \|x_1 - c_k\|^{-2}}$$

Donde $\sum_{k=1}^{2} \|x_1 - c_k\|^{-2} = \|x_1 - c_1\|^{-2} + \|x_1 - c_2\|^{-2} = 0.021$

Entonces:

$$u_{2,1} = \frac{0.001}{0.021} = 0.04$$

- Los demás elementos de la matriz se calculan de forma similar. La matriz queda como sigue (en esta iteración):

$$U = \begin{bmatrix} 0.95 & 0.55 & 0.57 & 0.46 \\ 0.04 & 0.45 & 0.43 & 0.54 \end{bmatrix}$$

Ejemplo

- Se calculan los centros de los grupos:

$$c_1 = \frac{0.95^2 \begin{bmatrix}10\\0\\0\end{bmatrix} + 0.55^2 \begin{bmatrix}10\\10\\0\end{bmatrix} + 0.57^2 \begin{bmatrix}9\\7\\1\end{bmatrix} + 0.46^2 \begin{bmatrix}3\\8\\5\end{bmatrix}}{0.95^2 + 0.55^2 + 0.57^2 + 0.46^2} = \begin{bmatrix}8.97\\4.1\\0.8\end{bmatrix}$$

$$c_2 = \frac{0.04^2 \begin{bmatrix}10\\0\\0\end{bmatrix} + 0.45^2 \begin{bmatrix}10\\10\\0\end{bmatrix} + 0.43^2 \begin{bmatrix}9\\7\\1\end{bmatrix} + 0.54^2 \begin{bmatrix}3\\8\\5\end{bmatrix}}{0.04^2 + 0.45^2 + 0.43^2 + 0.54^2} = \begin{bmatrix}6.73\\8.31\\2.42\end{bmatrix}$$

- Se vuelven a realizar todos los cálculos hasta minimizar la función de similitud.

Ejercicio

Se recomienda usar el conjunto de datos que provienen del Instituto de Oncología, para hacer un agrupamiento de datos relativos a observaciones de pacientes con cáncer:

Conjunto de datos: Incluye 201 instancias de una clase y 85 instancias de otra clase las instancias son descritas por 9 atributos, algunos son lineales y otros son nominales	Creators: Matjaz Zwitter & Milan Soklic (physicians) Institute of Oncology University Medical Center Ljubljana, Yugoslavia Donors: Ming Tan and Jeff Schlimmer (Jeffrey.Schlimmer **'@'** a.gp.cs.cmu.edu)

Attribute Information:

1. Class: no-recurrence-events, recurrence-events
2. age: 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70-79, 80-89, 90-99.
3. menopause: lt40, ge40, premeno.
4. tumor-size: 0-4, 5-9, 10-14, 15-19, 20-24, 25-29, 30-34, 35-39, 40-44, 45-49, 50-54, 55-59.
5. inv-nodes: 0-2, 3-5, 6-8, 9-11, 12-14, 15-17, 18-20, 21-23, 24-26, 27-29, 30-32, 33-35, 36-39.
6. node-caps: yes, no.
7. deg-malig: 1, 2, 3.
8. breast: left, right.
9. breast-quad: left-up, left-low, right-up, right-low, central.
10. irradiat: yes, no.

Lo datos obténgalos de

http://archive.ics.uci.edu/ml/datasets/Breast+Cancer

Referencias

- Janos Abonyi, Balazs Feil, “Cluster analysis for data mining and system identification”. Birkhauser Verlag, 2007.
- Jyh-Shing Roger Jang, Chuen-Tsai Sun, Eiji Mizutani, “Neuro-fuzzy and soft computing: a computational approach to learning and machine intelligence”. Prentice Hall, 1997.
- Shoichiro Nakamura, “Métodos numéricos aplicados a software”. Prentice Hall, 1992
- Pang -Ning Tan,, Michael Steinbach && Vipin Kumar:: Introduction Introduction to Data Data Mining Addison--Wesley, 2006.. ISBN 0321321367