TECNICAS DE LOGICA DIFUSA APLICADAS A LA MINERIA DE DATOS.

Fuzzy Logic Techniques for Data Mining

RESUMEN
La minería de datos permite manejar y clasificar grandes cantidades de datos, una de las tareas típicas de la minería de datos es el *clustering o agrupamiento*, Fuzzy C-Means es una técnica difusa de minería de datos para el *clustering* que se basa en el algoritmo clásico C-Means.

Fuzzy C-Means asigna a cada dato un grado de pertenencia dentro de cada cluster y como consecuencia un dato puede pertenecer parcialmente a más de un grupo. Fuzzy C-Means se utiliza en tareas de clasificación de datos, bioinformática, economía, entre otras.

PALABRAS CLAVES: *clustering*, C-Means, Fuzzy C-Means, Minería de datos, técnicas de lógica difusa.

ABSTRACT
Data mining lets handle and classify great amounts of data, one common task of data mining is clustering, Fuzzy C-Means is a data mining clustering fuzzy technique based on classic C-Means algorithm.

Fuzzy C-Means assigns a membership grade to each data inside each cluster and as a result a data cans partially belong in more than one group. Fuzzy C-Means is used in data classification, bioinformatics, economy and other knowledge areas.

KEYWORDS: *clustering, C-Means, data mining, Fuzzy C-Means, fuzzy logic techniques.*

JERÓNIMO ROJAS DIAZ
Ing. Electrónico
Estudiante de Maestría
Énfasis Ciencias Computacionales
Profesor Auxiliar
Universidad Tecnológica de Pereira
Profesor Universidad de Caldas
jeroxx@gmail.com

JULIO CESAR CHAVARRO PORRAS
Ph.D. (c) en Ingeniería. Área de énfasis: Ciencias de la computación
Profesor Asistente
Universidad Tecnológica de Pereira
jchavar@utp.edu.co

RICARDO MORENO LAVERDE
M.Sc. en Administración
Económica y Financiera
Profesor Asistente
Universidad Tecnológica de Pereira
rmoreno@utp.edu.co

1. INTRODUCCIÓN

La gran cantidad de datos y el elevado volumen de información que se tienen actualmente ha hecho necesario contar con técnicas automáticas que permitan indagar, organizar y extraer información implícita presente en las enormes bases de datos que contienen información bibliográfica, económica, genética, información derivada de experimentos e investigaciones y muchos otros tipos de información la cual extraer de forma manual resulta prácticamente imposible a medida que va creciendo el tamaño de las bases de datos.

Para manipular automáticamente bases de datos grandes surge la minería de datos que se puede definir como un proceso analítico diseñado para explorar grandes cantidades de datos con el objetivo de detectar patrones de comportamiento consistentes o relaciones entre las diferentes variables para aplicarlos a nuevos conjuntos de datos. [1] Ocasionalmente se hace referencia a la minería de datos como descubrimiento de conocimiento (*knowledge discovery*) otra definición de la minería de datos es la extracción no trivial de información implicita, previamente desconocida y potencialmente útil, a partir de los datos. [2]

Para realizar este proceso no trivial de identificación válida, novedosa, potencialmente útil y entendible de patrones comprensibles que se encuentran ocultos en los datos [3] se utilizan técnicas estadísticas que permiten reunir y analizar los datos, las técnicas estadísticas trabajan con la probabilidad o las distribuciones de probabilidad de los datos, así existen estimadores de probabilidad máxima que permiten encontrar clases diferentes en un conjunto de datos, clasificadores Bayesianos que permiten hacer inferencia, procesos o cadenas ocultas de Markov para encontrar patrones ocultos en los datos, entre otras técnicas.

La minería de datos también utiliza técnicas provenientes de la inteligencia artificial, así se pueden encontrar redes

Fecha de Recepción: (Letra Times New Roman de 8 puntos)
Fecha de Aceptación: Dejar en blanco

neuronales, algoritmos genéticos, técnicas de lógica difusa y otros tipos de heurísticas aplicados a los procesos de extracción de patrones ocultos o información útil de los conjuntos de datos.

Tanto las técnicas estadísticas como las técnicas de inteligencia artificial son bastante poderosas; en algunos casos, estas son tan solo dos enfoques o alternativas diferentes para la solución de un mismo problema, en otras ocasiones son técnicas complementarias porque resuelven problemas de naturaleza diferente.

Las técnicas de lógica difusa permiten manejar datos en los cuales existe una transición suave entre categorías distintas, por lo que algunos datos pueden tener propiedades de clases diferentes, estando parcialmente en más de un grupo con un grado específico de pertenencia.

Una de las tareas de la minería de datos es la identificación de *grupos o clusters naturales* en los conjuntos de datos. Una técnica difusa bastante conocida y que ha cobrado importancia en la tarea de *clustering o agrupamiento* es el algoritmo Fuzzy C-Means (FCM) el cuál es una extensión difusa del conocido C-Means.

2. C-MEANS *CLASICO*

C-means es un algoritmo iterativo que hace parte de las técnicas de *agrupamiento* no supervisado y tiene como objetivo encontrar patrones o grupos interesantes en un conjunto de datos dado, de tal manera que tales patrones, estructuras o grupos encontrados sirvan para clasificación, diseño de estrategias, soporte de decisiones, organización de la información, entre otras.

C-means al igual que otras técnicas clásicas de *agrupamiento* realiza una *partición dura* del conjunto de datos, tal partición se caracteriza porque cada dato pertenece exclusivamente a un *cluster* (grupo o clase) de la partición, además, los *clusters* deben cubrir totalmente el conjunto de datos, es decir cada dato tiene que pertenecer a alguno de los *clusters*; la cantidad de clusters debe ser definida para inicializar el algoritmo. Una partición dura se define formalmente como sigue:

Sea X un conjunto de datos y x_i un elemento perteneciente a X. se dice que una partición $P = \{C_1, C_2, ..., C_c\}$ donde c es un número entero no negativo que indica la cantidad de clusters, es una *partición dura* de X si y solo si:

1) $\forall x_i \in X \quad \exists C_j \in P$ tal que $x_i \in C_j$

2) $\forall x_i \in X \quad x_i \in C_j \Rightarrow x_i \notin C_k$

Donde $k \neq j, C_k, C_j, \in P$

La primera condición asegura que la partición cubra todos los puntos de X, la segunda garantiza que todos los clusters sean mutuamente excluyentes. [4]

Los objetivos del algoritmo c-means convencional son: encontrar el centro de cada cluster (este punto central es conocido con el nombre de *prototipo* del cluster) y determinar cual es el único cluster al que pertenece cada punto del conjunto de datos.

Para lograr el objetivo de hallar el centro de cada cluster se establece un criterio de búsqueda de dicho centro. Uno de tales criterios es la suma de la distancia entre los puntos de cada cluster y su centro, así:

3) $$J(P,V) = \sum_{j=1}^{c} \sum_{x_i \in X} \left\| x_i - v_j \right\|^2$$

Donde V es un vector de los centros de cada cluster a ser identificados. Este criterio es útil porque un conjunto de centros de los clusters adecuado o correcto brindará un valor mínimo de la función J. [4]

Como primer paso el algoritmo c-means calcula la partición actual con base a los prototipos actuales, como segundo paso modifica los prototipos actuales usando un método de optimización (Ej. Gradiente óptimo) para minimizar la función J, luego estos dos pasos se repiten iterativamente hasta alcanzar algún criterio de parada que usualmente es la diferencia de los prototipos entre dos ciclos consecutivos; cuando el algoritmo alcanza su criterio de parada significa que la función J llegó a un mínimo local.

3. FUZZY C-MEANS

En muchas situaciones cotidianas ocurre el caso que un dato está lo suficientemente cerca de dos clusters de tal manera que es difícil etiquetarlo en uno o en otro, esto se debe a la relativa frecuencia con la cuál un dato particular presenta características pertenecientes a clusters distintos y como consecuencia no es fácilmente clasificado; fuzzy c-means (FCM) es un algoritmo que se desarrolló con el objetivo de solucionar tales inconvenientes.

El algoritmo FCM asigna a cada dato un valor de pertenencia dentro de cada cluster y por consiguiente un dato específico puede pertenecer parcialmente a más de un cluster. A diferencia del algoritmo c-means clásico que trabaja con una *partición dura*, FCM realiza una *partición suave* del conjunto de datos, en tal partición los datos pertenecen en algún grado a todos los clusters; una partición suave se define formalmente como sigue:

Sea X conjunto de datos y x_i un elemento perteneciente a X. se dice que una partición $P = \{C_1, C_2, ..., C_c\}$ es una

partición suave de X si y solo si las siguientes condiciones se cumplen:

4) $\forall x_i \in X \quad \forall C_j \in P \quad 0 \le \mu_{C_j}(x_i) \le 1$

5) $\forall x_i \in X \quad \exists C_j \in P$ tal que $\mu_{C_j}(x_i) > 0$

Donde $\mu_{C_j}(x_i)$ denota el grado en el cuál x_i pertenece al cluster C_j. [4]

Un tipo de *partición suave* especial es aquella en la que la suma de los grados de pertenencia de un punto específico en todos los clusters es igual a 1.

6) $\sum_j \mu_{C_j}(x_i) = 1 \quad \forall x_i \in X$

Una *partición suave* que cumple esta condición adicional es llamada una *partición suave restringida*. El algoritmo FCM produce una *partición suave restringida* y para hacer esto la función objetivo J se extiende de dos maneras, por un lado en la ecuación (3) se incorporan los grados de pertenencia difusos de cada dato en cada cluster, por otro lado se introduce un parámetro adicional m que sirve de peso exponente en la función de pertenencia, así la función objetivo extendida J_m es:

7) $$J_m(P,V) = \sum_{i=1}^{k} \sum_{x_k \in X} (\mu_{C_i}(x_k))^m \left\| x_k - v_i \right\|^2$$

Donde P es una partición difusa del conjunto de datos X formada por $C_1, C_2, \ldots C_k$. el parámetro m es un peso que determina el grado en el cuál los miembros parciales de un cluster afectan el resultado. [4] [5]

Al igual que c-means clásico, FCM también intenta encontrar una buena partición mediante la búsqueda de los prototipos v_i que minimicen la función objetivo J_m y adicionalmente, FCM también debe buscar las funciones de pertenencia μ_{C_i} que minimicen a J_m. Estas condiciones se presentan en el siguiente teorema que sirve como fundamento del algoritmo FCM:

3.1 Teorema Fuzzy C-Means

Una partición difusa $\{C_1, C_2, \ldots, C_k\}$ puede ser un mínimo local de la función objetivo J_m solo si las siguientes condiciones se cumplen [4] [5]:

8) $$\mu_{C_i}(x) = \frac{1}{\sum_{j=1}^{k} \left(\frac{\left\| x - v_i \right\|^2}{\left\| x - v_j \right\|^2} \right)^{\frac{1}{m-1}}} \quad 1 \le i \le k, x \in X$$

9) $$v_i = \frac{\sum_{x \in X} (\mu_{C_i}(x))^m \, x}{\sum_{x \in X}^{n} (\mu_{C_i}(x))^m} \quad 1 \le i \le k$$

Después de determinar el número de clusters, el valor de m, y el criterio de parada el algoritmo FCM efectúa dos pasos, primero calcula las funciones de pertenencia mediante la ecuación (8), como segundo paso actualiza los prototipos usando la ecuación (9), los dos pasos se repiten iterativamente hasta alcanzar el criterio de parada.

4. VALIDEZ DEL CLUSTER

La validez de un algoritmo de *agrupamiento (clustering)* se estima mediante un criterio objetivo para determinar que tan buena es la partición generada por el algoritmo. Estos criterios son importantes porque permiten comparar los resultados de diversos algoritmos y permiten determinar el mejor número de clusters.

Las medidas de la validez de una *partición suave restringida* son de tres categorías [4]:

a. Medidas basadas en el grado de pertenencia

Estas medidas calculan ciertas propiedades de las funciones de pertenencia en una *partición suave restringida,* una de estas medidas es el coeficiente de partición introducido por Jim Bezdek en 1973, este coeficiente mide el grado de *fuzzividad* (fuzziness) del cluster. Entre más difusos son los clusters, peor es la partición. La formula para esta medida es:

10) $V_{PC} = \frac{1}{n} \sum_{i=1}^{c} \sum_{j=1}^{n} \mu_{C_i}(x_j)$

El coeficiente de partición es igual a 1 en el caso que la partición sea *dura* y en el caso que la partición sea la más difusa posible es igual a 1/c. por tanto es más deseable que el coeficiente de la partición este lo más cercano a uno posible y de esta manera tener una buena partición. Las medidas de validez basadas en el grado de pertenencia como el coeficiente de partición pueden ser usadas para comparar particiones alternativas con igual número de clusters pero tienen la limitación de mejorar a medida que se aumenta la cantidad de clusters, tampoco consideran propiedades geométricas de la partición como la separación entre clusters, por estos motivos se desarrollaron las medidas geométricas. [4]

b. Medidas basadas en la geometría

Intuitivamente entre más compactos y separados estén los clusters mejor es una partición, una medida que considera

1. Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

estos aspectos fue introducida por X. Xie y G. Beni, su formula es:

$$11)\ V_{XB} = \left(\frac{\sum \sigma_i}{n}\right)\frac{1}{d_{\min}^2}$$

Donde σ_i es la variación del cluster C_i definida como:

$$12)\ \sigma_i = \sum_j \mu_{Ci}(x_j)\left\|x_j - v_i\right\|^2$$

n es la cardinalidad del conjunto de datos y d_{min} es la distancia más corta entre centros de clusters definida como:

$$13)\ d_{\min} = \min_{\substack{i,j \\ i \neq j}} \left\|v_i - v_j\right\|$$

El primer término de la ecuación (11) mide si un cluster no es compacto y el segundo término es una medida de la no separación entre clusters. Por lo tanto el producto de los dos refleja el grado en el cuál los clusters en una *partición suave* no son compactos y no están bien separados. Obviamente entre más pequeña es V_{XB} mejor es la partición. [4]

c. Medidas basadas en el desempeño

Estas medidas evalúan una partición con base a su desempeño respecto a un objetivo predefinido, por ejemplo el mínimo error en una prueba de clasificación.

5. CAMPOS DE APLICACIÓN

El algoritmo FCM es una técnica de minería de datos que permite encontrar grupos naturales en un conjunto de datos y puede ser aplicado en diversos campos como organización y clasificación de datos, reconocimiento de patrones, estudio del clima, diagnostico de enfermedades, bioinformática, genética, [6] cancelación de ruido e interferencia de una señal, estudio de series de tiempo, estudio de la rentabilidad económica de una empresa, [7] soporte de la decisión, segmentación de mercados y clientes (weber), entre otras.[8],[9],[10],[11]

6. EJEMPLO DE APLICACIÓN

Se tiene una tabla (tabla 1) en la cual se evalúa la rapidez y la resistencia de 11 futbolistas. Un valor cercano a 1 indica que el jugador es bastante rápido o resistente según el caso, un valor cercano a 0 muestra que el jugador es lento o poco resistente. Se desea separar el conjunto de datos en 2 grupos (cluster) para ver si se encuentran jugadores con características especiales, para esto se hará uso del algoritmo FCM.

La cantidad de cluster es 2, el parámetro m escogido es 2, los prototipos iniciales se definen inicialmente en v_1 = (0.2, 0.5) v_2 = (0.8, 0.5) y el criterio de parada para este ejemplo no se va tener en cuenta por que solo se va a realizar la primera iteración del algoritmo.

JUGADOR	RAPIDEZ	RESISTENCIA
1	0.58	0.33
2	0.90	0.11
3	0.68	0.17
4	0.11	0.44
5	0.47	0.81
6	0.24	0.83
7	0.09	0.18
8	0.82	0.11
9	0.65	0.50
10	0.09	0.63
11	0.98	0.24

Tabla 1. Evaluación de la rapidez y resistencia de 11 futbolistas.

Las funciones de pertenencia iniciales de ambos clusters son calculadas usando la ecuación (8):

$$\mu_{C1}(x_1) = \frac{1}{\sum_{j=1}^{2}\left(\frac{\left\|x_1 - v_1\right\|^2}{\left\|x_1 - v_j\right\|^2}\right)^2}$$

$$\left\|x_1 - v_1\right\|^2 = 0.38^2 + 0.17^2 = 0.1444 + 0.0289 = 0.1733$$

$$\left\|x_1 - v_2\right\|^2 = 0.22^2 + 0.17^2 = 0.0484 + 0.0289 = 0.0773$$

$$\mu_{C1}(x_1) = \frac{1}{\frac{0.1733}{0.1733} + \frac{0.1733}{0.0773}} = \frac{1}{1 + 2.2419} = 0.3085$$

De manera similar se obtienen los valores de las demás funciones de pertenencia, los resultados se muestran en la tabla 2.

Un grado de pertenencia de 1 indica la máxima pertenencia, mientras que un valor de 0 indica que el dato no pertenece al cluster, así los datos con mayor pertenencia al cluster 1 son el cuarto, quinto, sexto, séptimo y décimo, mientras que los datos restantes tienen una mayor pertenencia al cluster 2.

Dato	Pertenencia	
	Cluster1	Cluster2
1	0.3085	0.6915
2	0.2016	0.7984
3	0.2665	0.7335
4	0.9762	0.0238
5	0.5481	0.4519
6	0.7927	0.2073
7	0.8412	0.1588
8	0.2213	0.7787
9	0.1000	0.9000
10	0.9473	0.0527
11	0.1289	0.8711

Tabla 2. Valor de pertenencia de cada dato en cada cluster.

El siguiente paso del algoritmo es actualizar los prototipos según la ecuación (9)

$$v_1 = \frac{\sum_{k=1}^{11} (\mu_{C1}(x_k))^2 \; x_k}{\sum_{k=1}^{11} (\mu_{C1}(x_k))^2} = (0.1980, 0.5125)$$

$$v_2 = \frac{\sum_{k=1}^{11} (\mu_{C2}(x_k))^2 \; x_k}{\sum_{k=1}^{11} (\mu_{C2}(x_k))^2} = (0.7528, 0.2886)$$

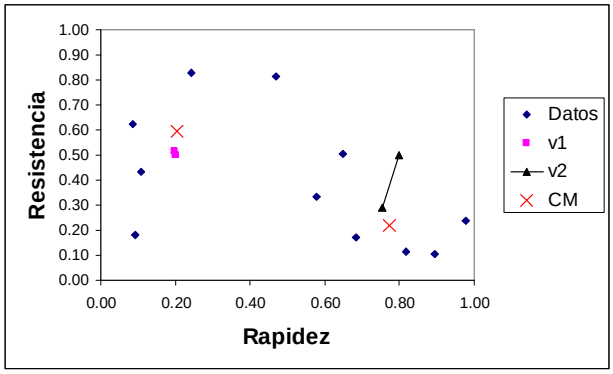

Figura 1. Dispersión de datos y centros de clusters.

La figura 1 muestra que el prototipo v_1 en la primera iteración se altera poco, mientras que las coordenadas del prototipo v_2 cambian considerablemente; con la ayuda de Matlab se realizaron 12 iteraciones y los prototipos resultantes fueron $v_1 = (0.2060, 0.5939)$ y $v_2 = (0.7729, 0.2192)$, se puede decir que el segundo cluster reúne los futbolistas que son rápidos pero poco resistentes, mientras que en el primer cluster están los jugadores no tan veloces pero más resistentes.

7. CONCLUSIONES Y RECOMENDACIONES

La minería de datos permite analizar, encontrar patrones, asociar, clasificar y agrupar grandes volúmenes de datos y para ello utiliza técnicas estadísticas o algoritmos provenientes de la inteligencia artificial.

Una de las tareas típicas de la minería de datos es el *clustering* de un conjunto de datos de entrada; uno de los algoritmos más conocidos para realizar *clustering* es C-Means pero este algoritmo produce una *partición dura* del conjunto de datos incluso en situaciones en las que un dato tiene características de grupos diferentes.

Fuzzy C-Means (FCM) es una extensión difusa de C-Means en la cuál un dato puede pertenecer parcialmente a más de un cluster por esta razón FCM, además, de calcular los prototipos del cluster, también calcula las funciones de pertenencia de los datos dentro de cada cluster.

FCM produce una *partición suave restringida* del conjunto de datos y por esto es útil en situaciones en las que los datos poseen características de distintos grupos.

Existen medidas de la validez de la partición, estas medidas se basan en características de la función de pertenencia, en la geometría de cada cluster o en el desempeño de la partición.

1. Las notas de pie de página deberán estar en la página donde se citan. Letra Times New Roman de 8 puntos

FCM puede ser aplicado en muchos campos como: clasificación de datos, medicina, bioinformática, economía, entre otras.

8. BIBLIOGRAFÍA

[1] Marín Llanes Luis A., Carro Cartaya Juan Carlos "La Minería de Datos como Herramienta en el Proceso de Inteligencia Competitiva" Consultoría Biomundi, Dirección de Inteligencia Corporativa, Instituto de Información Científica y Tecnológica (IDICT), CUBA, Taller Nacional sobre Inteligencia Empresarial IntEmpres'2000 [Online] available: http://www.redciencia.cu/empres/index.htm

[2] Frawley William J., Piatetsky-Shapiro Gregory, Matheus Cristopher J, "Knowledge Discovery in Databases: An Overview" pages 1--27. AAAI/MIT Press, 1991.

[3] Fayyad Usama, Piatetsky-Shapiro Gregory, Smyth Padhraic, "From Data Mining to Knowledge Discovery in Databases" American Association for Artificial Intelligence AAAI. Copyright © 1996

[4] Yen John, Langari Reza *Fuzzy Logic intelligence control and information*, New Jersey: Prentice Hall, 1999, p. 351-362.

[5] Klir George J, Yuan Bo *Fuzzy Sets and Fuzzy Logic theory and applications*, New Jersey: Prentice Hall, 1995, p. 357-362.

[6] Young Kim Seo, Won Lee Jae, Bae Jong Sung, "Effect of data normalization on fuzzy clustering of DNA microarray data" BMC informatics. 7:134 2006.

[7] Díaz Diez Bárbara, Morillas Raya Antonio, "Minería de Datos y Lógica Difusa. Una aplicación al estudio de la rentabilidad económica de las empresas agroalimentarias en Andalucía" Estadística Española. Vol. 46, Núm. 157, 2004 p. 409-430.

[8] Weber Richard, "Data Mining en la Empresa Y en las Finanzas Utilizando Tecnologías Inteligentes" *Revista Ingeniería de Sistema, Departamento de Ingeniería Industrial, Universidad de Chile, vol 14, Núm. 1, 2000*

[9] Martínez de Pisón Ascacibar F. Javier, "Optimización Mediante Técnicas de Minería de Datos del Ciclo de recocido de una línea de galvanizado" Tesis Doctoral, Universidad de la Rioja, 2003.

[10] Jantzen Jan, "Neurofuzzy Modelling" Technical University of Denmark, Department of Automation Aerospace Corp., Los Angeles, CA, Tech. report no 98-H-874 (nfmod), 30 Oct 1998

[11] Zha Wei, Chan Wei-Yip, "Objective Speech Quality Measurement Using Statistical Data Mining" EURASIP Journal on Applied Signal Processing 2005:9, 1410–1424.