Artículo **Revista de Análisis Cuantitativo y Estadístico**
Junio, 2016 Vol.3 No.7 26-29

Estudio de métodos difusos para la agrupación en un conjunto de datos

FUENTES, Juan Jaime*†

Universidad Politécnica del Valle del Évora, Angostura, Sinaloa, Mexico

Recibido Septiembre 28, 2015; Aceptado Enero 5, 2016

Resumen

En este documento se analiza los resultados de una investigación que realiza una comparación del agrupamiento de las caracteristicas fisicas de semillas de frijol, usando dos métodos de lógica difusa; el primer caso utiliza la lógica difusa con la Fuzzy C-Means algoritmo (FCM), que trata de encontrar similitudes entre las diferentes variables.

Lógica Difusa, Fuzzy C-Means (FM), Sistema de Inferencia Difusa, Agrupación, Frijol

Abstract

This research explain the results in investigation of comparison of a glustering of physical characteristics bean seeds, using two methods of fuzzy logic; the first case uses fuzzy logic with Fuzzy C-Means algorithm (FCM), trying to find similarities between different variables.

Fuzzy Logic, Fuzzy C-Means (FM), Fuzzy Inference System (FIS), Clusters, Beans

Citación: FUENTES, Juan Jaime. Estudio de métodos difusos para la agrupación en un conjunto de datos. Revista de Análisis Cuantitativo y Estadístico. 2016. 3-7: 26-29

*Correspondencia al Autor (Correo Electrónico: juanjaime.fuentes@upve.edu.mx)
† Investigador contribuyendo como primer autor.

Introducción

La agrupación de características es utilizada cuando no existe el conocimiento suficiente sobre los valores de las caracteisticas de un objeto de estudio. Esto puede ser extendido a la gran mayoria de las ramas deinvestigación Biotecnología, Medicina, Ciencias Computacionales, Ciencias Administrativas, etc. El principal objetivo de interactuar con estos datos es la clasifican en pequeños grupos que describan sus características principales, basándose en la similitud o diferencia entre ellos. Es imposible analizar directamente dicha cantidad de datos, por lo que comúnmente se recurre a la utilización de técnicas de agrupación que ayudan a particionar un conjunto de datos en pequeños grupos que permiten un análisis eficiente de la información. El análisis de grandes volúmenes de datos no sólo puede brindar información adicional, sino también conocimiento nuevo.

Marco teórico

Esencialmente, en un proceso que conduzca a la solución de un problema, este busca analizar los datos disponibles. El análisis de los datos forma el núcleo de la minería de datos, pero el proceso completo abarca también temas tales como la definición del problema y el desarrollo del problema para resolverlo.

El descubrimiento de esta información oculta es posible gracias a la Minería de Datos (del inglés Data Mining), que entre otras sofisticadas técnicas aplica la inteligencia artificial para encontrar patrones y relaciones dentro de los datos permitiendo la creación de modelos, es decir, representaciones abstractas de la realidad, pero es el descubrimiento del conocimiento, KDD, por sus siglas en inglés. (Morrison, 2014) el que se encarga de la preparación de los datos y la interpretación de los resultados obtenidos, los cuales dan un significado a estos patrones encontrados.

Las técnicas de minería de datos son el resultado de un largo proceso de investigación y desarrollo de productos. Esta evolución comenzó cuando los datos fueron almacenados por primera vez en computadoras, y continuó con mejoras en el acceso a los datos, y más recientemente con tecnologías generadas para permitir a los usuarios navegar a través de los datos en tiempo real. La minería de datos toma este proceso de evolución más allá del acceso y navegación retrospectiva de los datos, hacia la entrega de información prospectiva y proactiva. La minería de datos está lista para su aplicación en la comunidad científica porque está basado en tres tecnologías que ya están suficientemente maduras en computación (Zaïane, 2007):

- Recolección masiva de datos.
- Potentes computadoras con multiprocesadores.
- Algoritmos de minería de datos

Para las empresas, el aprovechamiento de la minería para datos de estas características representa retos tanto en infraestructura de almacenamiento y procesamiento como en la captación de personal capacitado que pueda adaptar e innovar para las aplicaciones especıficas. Solamente en Estados Unidos de América, se estima que en el año 2018 habrá una necesidad de 140 000 a 190 000 expertos con estos conocimientos (Stackpole, 2012).

Existen algoritmos que apoyan a la creación de agrupaciones y son los de índoles difusas los que más éxito han tenido como lo son: El algoritmo Fuzzy C-Means(FCM) Dumm(1973) pertenece a una clase de algoritmos basados en funciones objetivo, en cambio el algoritmo K –means propuesto por Chih et al. (2011) es un método de agrupamiento, que tiene como objetivo la partición n datos en k grupos en el que cada grupo pertenece al grupo más cercano a la media.

ISSN 2410-3438

Desarrollo

Se realizó un estudio para evaluar el comportamiento del algoritmo Fuzzy C-Means(FCM), Fuzzy K-Means(FKM) en la clasificación de semillas de frijol azufrado higuera. La ejecución de los métodos de agrupación difusa se realizó tomando conjunto de datos reales obtenidos de manera propia a través de las características propuestas por Celis (2008) la cual se presenta en la tabla 1.

Variable
Índice cotiledón/testa índice
Cotiledón/eje embrionario
Intensidad (del color)
Luminosidad (del color)
Peso volumétrico Peso de 100 Semillas
Porcentaje de cotiledón
Porcentaje de testa
Porcentaje de eje embrionario
Tono (del color)

Tabla 1 Características generales de una semilla de frijol azufrado higuera

Los cuales son utilizados para el entrenamiento de los algoritmos, estos datos constan de 100 semillas de frijol azufrado higuera de la cuales se clasifican en dos tipos óptimo y no óptimo teniendo como resultado muestra 52 semillas con un resultado óptimo y 48 no óptimo, para obtener la partición de dos grupos.

Se compararon los resultados con respecto a los mismos experimentos. La evaluación de los resultados obtenidos en el estudio de los diferentes métodos difusos; se evalúa de la siguiente manera. Suponiendo que el número final de clasificación es k, la medida de exactitud r está dada por la ecuación siguiente:

$$r = \frac{\sum_{i=1}^{k} a_i}{n} \quad (1)$$

Donde n es el número de instancias del conjunto de datos, ai es el número de instancias que aparecen clasificadas correctamente en i y en su correspondiente clase, la cual es aquella que tenga el número máximo. En otras palabras, ai es el número de instancias con las etiquetas de la clase que dominan en la clasificación i. Por lo tanto, la prueba del algoritmo se realizó con diferentes números de atributos, en los cuales cada caso se evalúa con la ecuación siguiente; para el desarrollo de estos algoritmos fue utilizada la herramienta computacional de Matlab®.

$$E = 1 - r \quad (2)$$

Implementación y resultados

Se realizó un número de entrenamientos que permitieron obtener diferentes resultados, siendo el conjunto de datos utilizados para el algoritmo FCM el mejor resultado fue obtenido con el entrenamiento de las variables Índice Cotiledón, Luminosidad, Peso Volumétrico, Tono. Con el que se obtiene un Error del 0.02, que en números porcentuales estamos diciendo que tenemos un 99,98% de clasificación correcta de los 100 datos entrenados los cuales se muestran en la tabla 2.

Agrupación	Optimo	No optimo	Instancias
1	1	47	48
2	51	1	52
Total	52	48	100
Error =	0.02		

Tabla 2 Entrenamiento realizado con los Datos Índice Cotiledón, Luminosidad, Peso Volumétrico, Tono Algoritmo Fuzzy C-Means.

Para el entrenamiento del Algoritmos FKM del conjunto de los 100 datos utilizados, al igual que con el algoritmo anterior el mejor resultado obtenido fue utilizando las variables Índice Cotiledón, Luminosidad, Peso Volumétrico, Tono.

ISSN 2410-3438

El error que presenta este algoritmo es del 0.84, que en números porcentuales estamos diciendo que tenemos un 99,16% de clasificación correcta como se muestran en la tabla 3.

Agrupación	Optimo	No optimo	Instancias
1	8	48	48
2	44	8	52
Total	52	48	100
Error =	0.84		

Tabla 3 Entrenamiento realizado con los Datos Índice Cotiledón, Luminosidad, Peso Volumétrico, Tono Algoritmo Fuzzy K-Means

Conclusiones

Los resultados del análisis para el conjunto de datos de las semillas de frijol azufrado higuera, De manera individual lo obtuvo el algoritmo (FCM), de todos los entrenamientos el que obtiene el mejor error promedio para este conjunto de datos es el algoritmo (FKM) como se muestra en el Tabla 4. Los algoritmos propuestos presentan buen desempeño para los conjuntos de datos evaluados en este trabajo. En general, los algoritmos en la mayoría de las pruebas evaluadas, presentan un comportamiento estable en la medición del error. Por lo tanto, esto implica que los resultados en los procesos de agrupamiento mantienenestable la variabilidad de los datos dentro de los conjuntos de datos.

Método	Error	%	Error Promedio
FCM	0.02	99.98%	0.18434
FKM	0.84	99.16%	0.11326

Tabla 4 Resultados obtenidos del procesamiento de datos con los Algoritmos Fuzzy C-Means, Fuzzy K-Means

Recomendaciones

Se puede mejorar el mecanismo para el manejo de datos anómalos en el conjunto de datos, de forma que se encuentren nuevas soluciones para asignar o eliminar estos datos de los agrupamientos.

Otra recomendación que se puede hacer es la creación de algún algoritmo de clasificación difusa que pueda trabajar con datos mixtos y no solo con numéricos de esta manera incrementar la complejidad de la búsqueda de nueva información.

Referencias

Celis Velázquez, R. Características morfológicas y fisiológicas de la semilla de frijol (Phaseolus vulgaris L.) domesticado y silvestre y su relación con el desarrollo y establecimiento de la plántula, (Tesis doctoral inédita), Colegio de Postgraduados, 2008

Chih T.C; Jim Z. C; Mu-der J., A Fuzzy K-means Clustering Algorithm Using Cluster Center Displacement, journal of information science and engineering No. 27, 2011, 995-1009.

Dunn, J. C. A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact WellSeparated Clusters, Cybernetics and Systems,Vol 3 No. 3, 1973, 32-57

Morrison, D. (2014). Phylogenetic networks: a new form of multivariate data summary for data mining and exploratory data analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 4(4), pp.296-312.

Stackpole, B. Your Big Data To-Do List. Computer World, Feb. 13 2012.

Yang, M.-S. A Survey of Fuzzy Clustering. Mathematical and Computer Modelling, Vol.18, No 11, 1993, 1-16.

Zaïane, O. Principles of Knowledge Discovery in Databases. University of Alberta. Department of Computing Science. 2007

ISSN 2410-3438