Artículo **Revista Tecnología e Innovación**
Diciembre 2015 Vol.2 No.5 929-941

El algoritmo de agrupamiento K-Modas: Un caso de estudio

RENDÓN, Eréndira†, ZEPEDA, Ricardo, BARRUETA, Elizabeth y ITZEL-MARÍA, Abundez

Departamento de Sistema y Computación, Instituto Tecnológico de Toluca

Recibido 5 de Julio, 2015; Aceptado 24 de Noviembre, 2015

Resumen

En este trabajo se desarrolló un software que utiliza el algoritmo K- modas para realizar agrupamiento con bases de datos descritas en datos categóricos, para probar el software se presenta un caso estudio, donde se encontrarán las caracterizas de los estudiantes que terminaron su carrera con un título. Las pruebas se realizaron con una base de datos del Instituto Tecnológico de Toluca de la carrera de Ingeniería en Sistemas Computacionales.

Algoritmos de agrupamiento, algoritmo K-Modas, datos categóricos.

Abstract

In this paper we developed a software that uses K-modas algorithm in order to cluster with databases described as categorical data. To test the software we present a study case, where the K-modas algorithm was used in order to find the students' features that finished their carrier with a degree. We worked with a data base of Instituto Tecnológico de Toluca, from Computational System Engineering carrier.

Clustering Algorithm, K-Modas algorithm, categorical data.

Citación: RENDÓN, Eréndira, ZEPEDA, Ricardo, BARRUETA, Elizabeth y ITZEL-MARÍA, Abundez. El algoritmo de agrupamiento K-Modas: Un caso de estudio. Revista de Tecnología e Innovación 2015, 2-5: 929-941

† Investigador contribuyendo como primer autor.

Introducción

El descubrimiento del conocimiento en bases de datos (KDD) es el proceso global de búsqueda de nuevo conocimiento a partir de los datos almacenados en las bases de datos. Este proceso incluye: filtrado, procesamiento, transformación, técnicas de minería de datos, interpretación y validación del conocimiento extraído (Fayyad U.M., 1996), ver figura 1.

Figura 1 Proceso KDD

La minería de datos es un paso importante en el proceso KDD. La minería de datos tiene dos tareas principales: las predictivas y las descriptivas. En las tareas descriptivas existen varias técnicas, tales como el agrupamiento (clustering), sumarización, modelado de dependencias. El agrupamiento es una técnica muy utilizada en las tareas de minería de datos, por esta razón, ha sido ampliamente estudiado debido a la gran variedad de aplicaciones donde se puede trabajar esta técnica. Se puede encontrar en la literatura una gran variedad de algoritmos de agrupamiento (Kaufman L, 1989), los cuales pueden ser utilizados en función del tipo de datos que trabajen, es decir si la base de datos está descrita en datos de tipo numérico o categórico. El algoritmo K-Modas es un algoritmo de agrupamiento (Zhexue, 1998) que trabaja con datos categóricos. En esta investigación, la base de datos que se utilizó está descrita con este tipo de datos, dicha base de datos contiene la descripción de los estudiantes del Tecnológico de Toluca de la carrera de ingeniería en Sistemas Computacionales.

De esta manera en este trabajo se utilizó el algoritmo K-modas para encontrar las características de los buenos estudiantes, es decir aquellos que terminan titulados.
Tratando con estudiantes, existen ciertos factores que influyen con el rendimiento y éxito académico que pertenecen al grupo de datos categóricos. (Tinto, 1992) se postula que los estudiantes ingresan a la universidad con diversas habilidades y patrones de características personales, familiares y académicas, incluidas metas y predisposiciones iniciales para asistir a la universidad. Estas últimas se modifican y reformulan continuamente a través de una serie de interacciones entre el individuo, las estructuras y miembros de los sistemas sociales y académicos de la institución.

Así nuestra investigación se centró en desarrollar un software que utiliza el algoritmo de agrupamiento K-modas para determinar los factores o características que influyen en el éxito o no de un estudiante (obtención del título) en una base de datos de estudiantes de ingeniería en sistemas computacionales del Instituto Tecnológico de Toluca. Es importante resaltar que el software desarrollado puede trabajar con otros tipos de base de datos.

El resto de este trabajo se encuentra organizado de la siguiente manera en la sección 1 se decriben los trabajos relacionados con la solución que se presenta, en la sección 2 se describen algunas definiciones necesarias para un mejor entendimiento del algoritmo del algoritmo K-modas, así como la descripción de éste, en la sección 3 se proporciona la metodología que se utilizó para la programación del software "K-modas7", en la sección 4 se describen los resultados obtenidos, finalemnete en la última sección se presentan las conclusiones a las que se llegaron.

Trabajos relacionados

Dentro del sector educativo se encuentran diversos elementos que permiten identificar el rendimiento y éxito académico de los estudiantes. En la actualidad existe un significativo interés por el estudio de las variables relacionadas con el éxito académico y la manera en que se comportan los resultados que se generan a través de diferentes técnicas y métodos. Existen investigaciones que han sido realizadas por expertos en el tema, aportando conocimiento para mejorar y analizar estas variables o factores, donde establecen que las condiciones académicas, la adaptación a la institución, las estrategias de aprendizaje y la situación socioeconómica son algunos de los elementos decisivos en el éxito escolar. Algunas de las investigaciones que se han realizado al respecto son:

En (Navarro, 2003) se menciona que existen diversas variables que pueden identificarse de la siguiente forma, en relación con los individuos, una de ellas son las características que son susceptibles de modificarse a través del proceso educativo y aquellas que no pueden modificarse, como las características genéticas y las experiencias previas. También establece que siempre que se pretende encontrar el fracaso escolar se apunta hacia los programas de estudio, la falta de recursos de las instituciones y rara vez se piensa en el papel que los padres juegan.

Se realizó una investigación por parte de (Martínez, 2003) acerca del perfil de éxito de un estudiante de posgrado, donde se indica que la obtención del grado a nivel posgrado es baja y repercute tanto en el ámbito social como educativo. Las variables que se relacionan dentro del estudio son el nivel de conocimientos previos, una mayor capacidad intelectual, características psicológicas, hábitos académicos positivos y algunas otras variables, tiene como resultado un mayor éxito académico.

En (Gómez, 2003) se tiene como objetivo investigar las características motivacionales, cognitivas y autorreguladoras, así como las actividades de aprendizaje que llevan durante la carrera de Química en la Universidad Nacional Autónoma de México. En este estudio se observa que los aciertos, razonamientos, estrategias y concepciones alternativas han contribuido a perfeccionar las áreas sobre el proceso de aprendizaje y la identificación del éxito en los estudiantes. Otros autores coinciden en que los factores personales y académicos determinan si un estudiante es exitoso o no al final de su carrera profesional (Acosta, 2004).

En (Belvis, 2009) se desarrolló un estudio que pretende determinar cuáles son los factores que afectan al rendimiento académico de los estudiantes universitarios en España. Se realizó una encuesta a una muestra de estudiantes de siete Facultades de Educación españolas, con lo cual se detectaron los factores que más inciden en el éxito o fracaso del estudiante son: la situación laboral; la dedicación y motivación por los estudios; las becas de estudio; las condiciones de acceso a la titulación y la preparación académica previa, así como el rendimiento académico que se consigue en los primeros semestres de estudio en la universidad. En este estudio se analizan e interpretan los resultados obtenidos y se realizan propuestas para mejorar las intervenciones y los servicios de apoyo para estudiantes.

En (Gatica, 2010) se menciona que "Los estudios universitarios representan demandas, compromisos, metas de mayor dificultad y exigencia. Se ha observado en la Facultad de Medicina un alto índice de reprobación y abandono durante los 2 primeros años de la licenciatura, el cual disminuye de manera importante en el área clínica". Por tal motivo se propone analizar las variables que intervienen en el rendimiento y éxito académico durante los primeros años de la carrera.

Ya que durante este periodo puede estar definida la continuidad de los estudios universitarios. En este estudio se dividen las variables en factores académicos, personales y socioeconómicos, tomando en cuenta el éxito académico como la acreditación oportuna de las asignaturas, exámenes departamentales y una puntuación determinada durante los primeros 2 años de la carrera Médico Cirujano de la Facultad de Medicina de la UNAM en la Ciudad de México.

El éxito académico del estudiante de licenciatura proporciona ciertos beneficios a la sociedad por su contribución al desarrollo económico, cultural y social del país, que se manifiesta en la productividad de sus actividades docentes, de investigación y difusión de la cultura.

Definiciones preliminares

Algoritmo de agrupamiento

El objetivo de los algoritmos de agrupamiento es encontar particiones disjuntas de un conjunto de datos o base de datos, de tal manera que los objetos en el misno grupos sean lo más similares que los objetos de los otros grupos (Jain, 1988).

Descripción del algoritmo k-modas

El algoritmo k-modas (Zhexue., 1998), fue diseñado para agrupar grandes conjuntos de datos categóricos, y tiene como objetivo obtener las k modas que representan al conjunto

Dominios y atributos categóricos

Zhexue en (Zhexue., 1998), describe los datos categóricos como objetos descritos únicamente por atributos categóricos o como una versión simplificada de los objetos simbólicos definidos en (Godwa, 1992).

Considera a todos los atributos numéricos (cuantitativos) al categorizarlos y no considera los atributos categóricos que están contenidos por una combinación de valores determinados. Los objetos y atributos categóricos aceptados por el algoritmo k-modas son definidos en (Zhexue., 1998).

Suponga que $A_1, A_2, \ldots, A_m$ son los m atributos que describen a un objeto en un espacio Ω y dominio$DOM(A_1), DOM(A_2), \ldots, DOM(A_m)$. Un dominio $DOM(A_j)$ es definido como categórico si es un conjunto finito y no ordenado. Ω Es un espacio categórico si todo $A_1, A_2, \ldots, A_m$ es categórico.

Objetos categóricos

Como en (Godwa K.C., 1991), un objeto categórico $X \varepsilon\, \Omega$ es representado como la conjunción lógica de pares atributo-valor $[A_1 = X_1] \wedge [A_2 = X_2] \wedge \ldots \wedge [A_m = X_m]$, donde $X_j\, \varepsilon\, DOM(A_j)$, para 1 $1 \leq j \leq m$ mismo para atributo-valor $[A_j = X_j]$ es llamado selector. X es un vector de la forma $[X_1, X_2, \ldots, X_m]$ y cada objeto en Ω tiene exactamente m valores atributos y si el valor para el atributo A_j no está disponible para un objeto X , entonces $A_j = \varepsilon$ donde ε representa al valor de un atributo no disponible.

Sea $X = \{X_1, X_2, \ldots, X_n\}$ un conjunto de n objetos categóricos $X \subseteq \Omega$ .El objeto X_i es representado como $[X_{i1}, X_{i2}, \ldots, X_{im}]$. Dos objetos X_i, X_k son iguales $X_i = X_k$ si $x_{ij} = x_{kj}$ para todo$1 \leq j \leq m$. La relación $X_i = X_k$ no quiere decir que X_i, X_k sean algunos objetos en las bases de datos del mundo real. Esto implica que dos objetos tienen igual valor categórico en sus atributos $A_1, A_2, \ldots, A_m$.

Asuma que *X* consiste de *n* objetos en donde *p* objetos son distintos. Sea *N* la cardinalidad del producto cartesiano $DOM(A_1)x\ DOM(A_2)x\ \dots x\ DOM(A_m)$.
Tenemos que $p \leq N$. De cualquier modo, *n* puede ser tan grande como ***N.***

Medidas de disimilaridad utilizadas

Sean *X, Y* dos objetos categóricos descritos por *m* atributos categóricos. La medida de disimilaridad entre *X* y *Y* se define por el total de las no coincidencias de los atributos categóricos de los objetos. El número más pequeño de las diferencias significa que los objetos son similares (Zhexue., 1998).

Formalmente:

$$d(X,Y) = \sum_{j=1}^{m} \delta(X_j, Y_i) \qquad (1)$$

Donde:

$$\delta(X_j, Y_i) = \begin{cases} 0\ (x_j = y_i) \\ 1\ (x_j \neq y_i) \end{cases} \qquad (2)$$

d(X,Y) da igual importancia a cada categoría del atributo. Si se toma en cuenta las frecuencias de las categorías en el conjunto de datos, se define la medida de disimilaridad,

Como:

$$d_{x^2}(X,Y) = \sum_{j=1}^{m} \frac{n_{xj} + n_{yj}}{n_{xj} n_{yj}} \delta(X_j, Y_j) \qquad (3)$$

Donde n_{xj} y n_{yj} son el número de objetos en el conjunto de datos, que tienen las categorías x_j y y_j para el atributo *j*. Zhexue denomina a la ecuación 3, distancia *xi-cuadrada* y la propone para descubrir grupos de objetos con baja representación en la base de datos.

Modas de un conjunto

Sea X un conjunto de objetos descritos por atributos categóricos. Una moda de X es un vector $Q = [q_1, q_2, \dots., q_m] \in \Omega$ que minimiza a $D(Q,X) = \sum_{i=1}^{n} d(X_i, Q)$ donde $X = (X_1, X_1, \dots\dots, X_n,)$ y d pueden ser calculadas con la ecuación 2 o la ecución 3.

Función criterio

Suponga que $\{S_1, S_2, \dots, S_k\}$ es una partición de X donde $S_1 \neq \emptyset$ (conjunto vacío), para $1 \leq l \leq k$ y $\{Q_1, Q_2, \dots, Q_k,\}$ las modas de $\{S_1, S_2, \dots, S_k\}$. El costo total de la partición es definido por:

$$E = \sum_{l=1}^{k} . \sum_{i=1}^{n} y_{i,l} d(X_i, Q_l) \qquad (4)$$

Donde $y_{i,l}$ es un elemento de la matriz de la partición Y_{nxl} como en (Godwa, 1991) y *d* puede ser definida como la ecuación 1 o la ecuación 3. Similar al algoritmo *k-medias*, el objetivo de agrupar el conjunto *X* es encontrar un conjunto $\{Q_1, Q_2, \dots, Q_k,\}$ que puede minimizar E. La ecuación 4, puede ser minimizada por el algoritmo k-modas.

El algoritmo K-Modas

El algoritmo k-modas es una versión del k-medias para datos categóricos.

En k-modas se hacen 3 modificaciones a k-medias:

- Uso de diferentes medidas de disimilaridad.
- Sustitución de k medias por k modas para formar los centros.
- El método basado en las frecuencias de los datos para actualizar las modas.

La actualización de las modas se realiza en cada asignación de un objeto a su grupo, mientras que en k-medias es al final de cada iteración del algoritmo. El algoritmo k-modas al igual que el algoritmo k-medias produce soluciones óptimas locales, que dependen del conjunto de modas iniciales y el orden de los objetos en el conjunto de datos.

Descripción del algoritmo K-modas

Paso 1: Seleccionar k modas iniciales, una para cada grupo.

Paso 2: Asignar cada objeto a la moda más cercana utilizando la distancia d. Actualizar la moda del grupo después de cada asignación.

Paso 3: Después que todos los objetos han sido asignados a un grupo, volver a examinar la disimilaridad de los objetos con las modas actuales. Si un objeto es encontrado tal que su moda más cercana corresponde a otro grupo, asignar el objeto a su nueva moda y actualizar la moda de ambos grupos.

Paso 4: Repetir el paso 3 hasta que no existan objetos cambiados de grupo.
Metodología

La investigación es de tipo descriptiva y experimental, la cual consta de 3 etapas (descriptiva, iterativa y resultante), que representan la recolección y procesamiento de los datos, así como los resultados obtenidos.

Estas etapas se encuentran definidas a continuación.

Etapa descriptiva

La información se obtuvo a partir de la herramienta de análisis de documentos a través de las oficinas de Servicios Escolares y Desarrollo Académico del Instituto Tecnológico de Toluca.

Tomando como muestra a los alumnos de la carrera de Ingeniería en Sistemas Computacionales de las generaciones 2000 a 2003.

De acuerdo a los datos obtenidos se admitirán en el estudio a todos los alumnos que cumplan con los siguientes criterios:

- Contar con expediente individual en el Instituto Tecnológico de Toluca.
- Haber cursado la carrera sin ser provenientes de otra institución.
- Contar con la información completa de las variables estudiadas.

Las variables empleadas en el estudio han sido asignadas a partir de investigaciones dirigidas al análisis y comportamiento de los factores que influyen en el proceso académico del estudiante a nivel licenciatura, dichas investigaciones realizan procesos diferentes al momento de evaluar los factores, sin embargo, regularmente se encuentran dentro de una clasificación conformada por tres grupos:

1. Factores académicos.
2. Factores personales.
3. Factores socioeconómicos.

De acuerdo con la clasificación anterior se han elegido las variables que intervendrán de manera trascendental en el desarrollo del estudio, son definidas como variables independientes y señaladas a continuación:

a) Estado Civil.

b) Edad.

c) Trabajo.

d) Dependientes Económicos.

e) Institución de procedencia.

f) Tiempo de egreso.

g) Periodo de ingreso.

h) Promedio.

Se determinó como variable dependiente al éxito académico (obtención del título a nivel licenciatura) considerado 7 años a partir de la última generación evaluada.

Etapa iterativa

El paradigma de construcción de prototipos inicia con la comunicación. El ingeniero de software y el cliente encuentran y definen los objetivos globales para el software, identifican los requisitos conocidos y las áreas del esquema en donde es necesaria más definición. Entonces se plantea con rapidez una iteración de construcción de prototipos y se presenta el modelado (en la forma de un diseño rápido). El diseño rápido conduce a la construcción de un prototipo. Después, el prototipo lo evalúa el cliente/usuario y con la retroalimentación se refinan los requisitos del software que se desarrollará. (Pressman, 2005).

Siguiendo el modelo anterior, se plantearon diversos apartados para llevar a cabo la construcción de prototipos, validarlos y continuar con el desarrollo de la aplicación. A continuación se describen de manera práctica, dichos apartados.

Pantalla principal

De manera inicial se determinó el requerimiento de áreas de texto para visualizar los resultados.

Las opciones para elegir los parámetros de entrada (número de grupos a formar, tipo de ecuación y selección de modas iniciales). También contar con los botones para realizar las acciones de agrupamiento y las frecuencias de dominios. Así como la lógica principal del algoritmo de agrupamiento k-modas.

Posteriormente se identificó que se debería contar con ciertas validaciones, de acuerdo a las opciones elegidas como parámetros de entrada, ya que las variantes no son aplicables en todos los casos.

Se presentó el prototipo y se agregó la validación del número de grupos a formar, para que sea mayor o igual a 2, y menor al número total de elementos. A su vez se colocó la barra donde aparece el nombre y la ruta del archivo que se está utilizando para el agrupamiento.

Se colocó una barra de menú en la parte superior de la interfaz, originalmente con el apartado de “abrir” en la sección de archivo. Ya que con esta opción, se carga el archivo para ser analizado y agrupado.

Una vez identificado el agrupamiento de datos, se solicitó la creación de una rutina que permita guardar archivos de texto, con los resultados que genera la aplicación. Definiendo 3 tipos de archivos: 1. Resultados con etiqueta de grupo. 2. Resultados con etiqueta de grupo y los parámetros de entrada ocupados. 3. Resultados ordenados de acuerdo con las etiquetas de grupo.

De manera final se valoró y se integró la opción cerrar, para complementar el menú. Y comenzar con la asignación de teclas rápidas, así como el inicio de generar otros apartados dentro de la barra de menú.

ISSN-2410-3993

Barra de menú complementaria

Los demás apartados añadidos en la barra de menú (editar, herramientas, ayuda), se determinó mediante un cambio de color en editar, la creación de archivos a través de una consulta a la base de datos con la opción de herramientas y contar con una guía rápida e información del software.

Se redefinió la parte de generación de archivos, debido a que varía de acuerdo a los parámetros de la base de datos y las opciones que pueden desarrollarse al crear los archivos de texto, que serán utilizados para realizar el agrupamiento.

Se presentaron las diversas iniciativas y con los cambios requeridos, se validaron los apartados mencionados en los párrafos anteriores, para finiquitar el proceso en la creación de la aplicación. Tomando en cuenta que se encuentra abierta la posibilidad de futuras mejoras o modificaciones, en caso de ser requeridas.

Etapa resultante

Los datos que proporciona la aplicación k-modas7, serán representados en forma de grupos, etiquetando cada uno de sus elementos, para validar y determinar los factores que influyen en el desarrollo del estudiante para lograr la obtención del título y perfil de éxito académico.

Estos resultados podrán ser observados en la aplicación o también generar un archivo de texto, con los datos correspondientes. Los cuáles serán analizados por expertos del Departamento de Desarrollo Académico del Tecnológico de Toluca.

Finalmente en la Figura 2 pueden observarse las etapas del procesamiento de información en forma gráfica y simplificada.

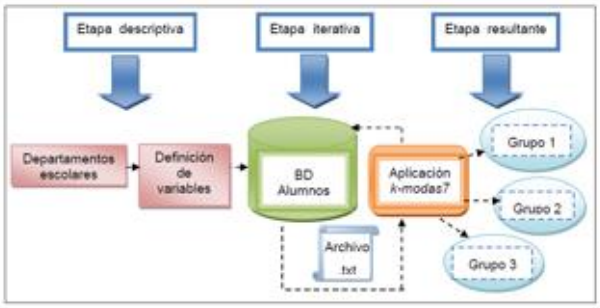

Figura 2 Etapas del procesamiento de información

La etapa descriptiva efectúa la recolección de datos y la definición de las variables que se utilizan en la investigación.

La parte intermedia se forma a partir de la base de datos y la interacción con la aplicación k-modas7 mediante archivos de texto.

Los resultados son identificados por círculos que representan agrupaciones de alumnos con características similares.

Aplicación K-Modas7

La aplicación k-modas7 es una herramienta que permite agrupar grandes cantidades de datos, mediante parámetros de entrada y archivos de texto. En primera instancia se presenta la pantalla inicial, para describir los elementos que la contienen, que puede observarse en la Figura 3.

Figura 3 Pantalla inicial

Existen 3 elementos principales en los que se compone la interfaz principal, hay una barra de menú en la parte superior, en la cual se efectúan las diversas acciones para iniciar el proceso de agrupamiento, así como opciones de edición y ayuda.

En el menú Archivo se elige el documento de texto (.txt) que se va agrupar, contando con una serie de datos identificados por un separador y con el mismo número de elementos por cada registro.

Si no se cuenta con un archivo de datos elaborado, se utiliza el menú Herramientas para generar un archivo, haciendo una consulta a la base de datos para obtener la información necesaria, para ser agrupada.

El menú Editar permite cambiar de color los datos resultantes en pantalla, y en el menú de Ayuda vienen una serie de instrucciones que sirven de apoyo para el uso de la aplicación. Para la segunda parte, puede verse una serie de opciones a elegir. En las que se encuentra el número de grupos a formar, el tipo de ecuación y la elección de las modas iniciales. Esta sección se representa por el número 2, se debe llenar el cuadro de texto con la cantidad de grupos que deseamos formar, posteriormente las ecuaciones con las que cuenta el algoritmo es la ecuación binaria y xi-cuadrada. Para finalizar la elección de parámetros, seleccionar entre primeros k elementos o modas ficticias.

En la última zona de la interfaz, se pueden ver los resultados que generan el archivo elegido, dominios de frecuencias y el agrupamiento. Todo de acuerdo a los parámetros seleccionados y que se mencionaron anteriormente.

A continuación se presenta una serie de pasos, para hacer uso correcto de la aplicación.
Ejecutar la aplicación k-modas7 para iniciar el proceso.

Seleccionar el menú Archivo-Abrir, ubicado en la Figura 4 y elegir un documento de texto almacenado en el equipo.

Figura 4 Barra de menú

Generar un archivo en el menú Herramientas, identificado en la Figura 5 a través de una consulta a la base de datos, para crear el documento de texto (.txt) que se estará agrupando con el uso de la aplicación. Hay que determinar los parámetros de conexión a la base de datos, así como la tabla y las condiciones necesarias para hacer uso de este apartado. Finalizando con el nombre del archivo a crear.

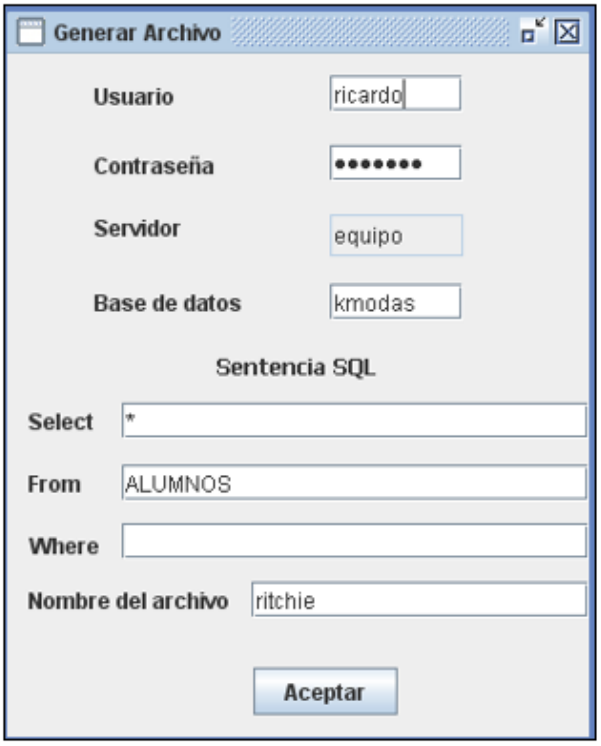

Figura 5 Pantalla generar archivo

Una vez seleccionado el archivo que se desea agrupar, se deberá escribir el número de grupos a formar (k), con el cual se determina las particiones con las que contaran los resultados finales.

Se debe tomar en cuenta que estos grupos deben ser mayores a 1 y menores al total de elementos para analizar.

Después se tiene que elegir el tipo de ecuación que utilizará el algoritmo k-modas, dentro de la aplicación, contando con las opciones de ecuación 1 (binaria) o ecuación 2 (xi-cuadrada).

En caso de seleccionar la ecuación 2, se deberá obtener las frecuencias de dominios para poder continuar el proceso, dando clic en el botón de Obtener frecuencias.

Ahora se tendrá que seleccionar el método de elección de modas iniciales, ya que deben crearse una serie de modas para que a partir de ellas se genere el agrupamiento. Se cuenta con las opciones de primeros k elementos y modas ficticias. A continuación se muestran los parámetros mencionados en la Figura 6.

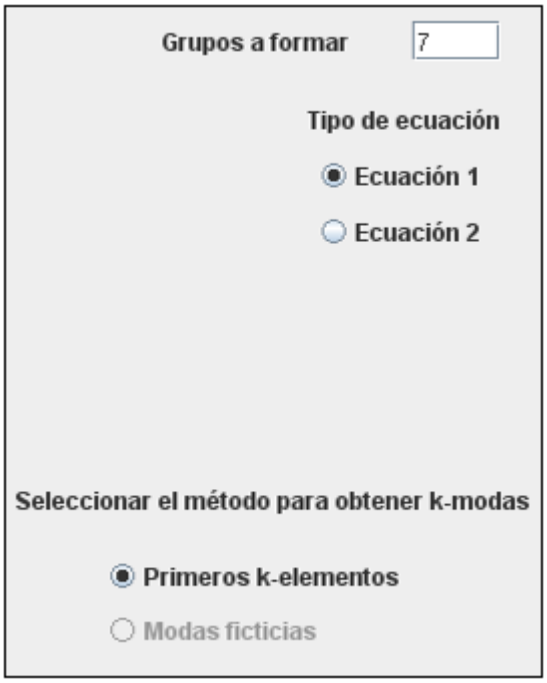

Figura 6 Sección de parámetros de entrada

Presionar el botón efectuar agrupamiento, para que se generen los resultados correspondientes, de acuerdo a los parámetros seleccionados. En esta parte termina el proceso que genera los grupos, para hacer uso de estos hay que hacer clic en el menú Archivo-Guardar, con lo que se va a generar una carpeta que contiene 3 archivos de texto para utilizar los resultados agrupados.

Resultados

Diseño de pruebas

Para desarrollar las pruebas de la investigación, fue necesario utilizar diversos parámetros que determinan el rumbo del proceso y de los resultados.

Se tendrá que elegir inicialmente un archivo de texto que contenga los datos para analizar, después se debe asignar el número de grupos a formar (k), seleccionar el tipo de ecuación (binaria o xi-cuadrada) y finalmente el método para determinar las modas iniciales (primeros k-elementos o modas ficticias).

De acuerdo a las opciones antes mencionadas, se presentará de manera estructurada, las posibles combinaciones para realizar las pruebas necesarias en el estudio, de acuerdo a la figura 7.

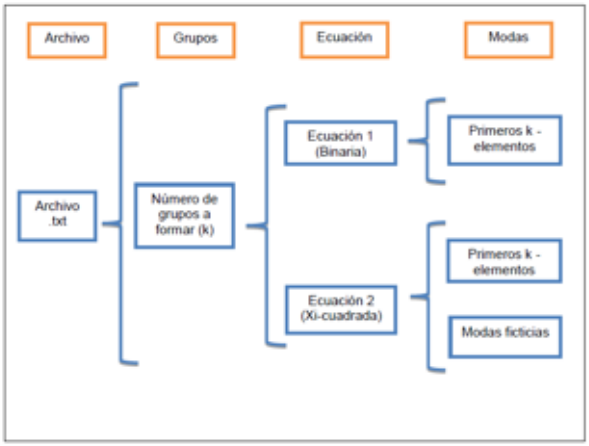

Figura 7 Representación estructurada del diseño de las pruebas.

Descripción de las pruebas realizadas

En esta sección se muestran los resultados obtenidos al agrupar el archivo alumnos_145.txt con la aplicación k-modas7, empleando diferentes opciones de configuración, al elegir el número de grupos a formar, el tipo de ecuación y la forma de elección de modas iniciales.

Para llevar a cabo las diversas pruebas, se utilizó principalmente un equipo Dell Inspiron 1420, con procesador Core2 Duo, una memoria RAM de 2GB y un disco duro de 80 GB.

Resultados de las pruebas

En este apartado se describen los resultados de las pruebas desarrolladas, presentando una parte de la tabla que contiene los parámetros utilizados en la aplicación, así como los datos finales del perfil del estudiante y la pureza de la agrupación. Ver Tabla 1.

Parámetros / Pruebas	Prueba 1		Prueba 2		Prueba 3	
Grupos a formar	k=2					
Tipo de ecuación	Ecuación 1		Ecuación 2			
Modas iniciales	Primeros k elementos		Primeros k elementos		Modas Ficticias	
G0	(S,27,SI,0,PREPARATORIA ESTATAL,4,5,AGOSTO - DICIEMBRE 2001,81,SI)	82.00%	NA	NA	(S,26,NO,0,PREPARATORIA ESTATAL,5 5,AGOSTO - DICIEMBRE 2003,84,SI)	60.42%
G1	(S,26,NO,0,PREPARATORIA ESTATAL,5 5,AGOSTO - DICIEMBRE 2003,84,NO)	51.58%	(S,26,NO,0,PREPARATORIA ESTATAL,5 5,AGOSTO - DICIEMBRE 2003,84,SI)	60.14%	NA	NA
G2						
G3						
G4						
G5						
G6						
G7						
G8						
G9						

Tabla 1 Resultados de las pruebas

Los resultados finales se generan con 15 pruebas y diferentes variantes en la elección de parámetros de entrada.

Para la aplicación que emplea el algoritmo de agrupamiento k-modas. En la tabla 1 se presentan los resultados de las tres mejores pruebas, las cuales obtuvieron la mejor pureza de grupos.

A partir de los experimentos desarrollados puede verse que en la prueba donde k=4, la ecuación empleada es la binaria (ecuación 1) y el método de elección de modas es primeros k elementos, que se encuentran tiene un porcentaje de los más altos en el estudio para los alumnos que si logran obtener el título de Ingeniería en Sistemas Computacionales (66 de 69 elementos), contando con una pureza de 95.65%. Para esta prueba el perfil de los estudiantes se describe con las siguientes características {Estado Civil = Soltero, Edad = 26, Trabajo = No, Dependientes Económicos = 0, Institución de procedencia = Preparatoria Estatal, Tiempo de Egreso = 5, Periodo de Ingreso =Agosto – Diciembre 2003, Promedio = 84, Titulo = Si}.

Para la prueba donde el número de grupos a formar (k)=5, ecuación 1 y primeros k elementos como modas iniciales, tiene una pureza del 100% para un grupo de 11 elementos. Esta prueba contiene las características principales de {Estado Civil = Soltero, Edad = 27, Trabajo = No, Dependientes Económicos = 0, Institución de procedencia = Preparatoria Estatal, Tiempo de Egreso = 6, Periodo de Ingreso = Agosto – Diciembre 2001, Promedio = 81, Titulo = Si}.

Como se puede ver el resultado en dos de las pruebas es igual, haciendo referencia a la moda planteada anteriormente {Estado Civil = Soltero, Edad = 27, Trabajo = No, Dependientes Económicos = 0, Institución de procedencia = Preparatoria Estatal, Tiempo de Egreso = 6, Periodo de Ingreso = Agosto – Diciembre 2001, Promedio =81, Titulo = Si}.

ISSN-2410-3993

Esto significa que el algoritmo, proporciona resultados confiables independientemente del número de grupos a formar (k), ya que se puede encontrar similitudes en los datos finales al momento de hacer diferentes pruebas, con diversos parámetros.

Conclusiones

Se diseñó un software que realiza agrupamiento de una base de datos con el algoritmo K-Modas. Además se conformó una base de datos con información de 145 alumnos pertenecientes a la carrera de Ingeniería en Sistemas Computacionales, tomando en cuenta las generaciones del año 2000 a 2003 en los diversos periodos escolares. Así la base de datos de pueba estuvo conformada por 150 registros descritos en los siguientes campos:

a) Estado Civil.

b) Edad.

c) Trabajo.

d) Dependientes Económicos.

e) Institución de procedencia.

f) Tiempo de egreso.

g) Periodo de ingreso.

h) Promedio.

Se utilizaron 2 medidas de disimilaridad en el estudio, la primera es la ecuación binaria y la segunda es la ecuación xi-cuadrada. El algoritmo k-modas fue evaluado para emplearlo en esta investigación, ya que puede ser utilizado con datos no numéricos.

De acuerdo a las 15 pruebas realizadas con el algoritmo de agrupamiento k-modas, pueden observarse diferentes resultados conforme a los parámetros de entrada que requiere para su funcionamiento. Los mejores resultados obtenidos encontraron que, aquellos estudiantes que podrán lograr la obtención del título de Ingeniería en Sistemas Computacionales deberán contar con las siguientes características:

Estado Civil: Soltero.
Edad: 27 años.
Trabajo: No.
Dependientes Económicos: 0.
Institución de Procedencia: Preparatoria Estatal.
Tiempo de Egreso: 6 años.
Periodo de Ingreso: Agosto – Diciembre 2001.
Promedio: 81.
Título: Si.

Referencias

Acosta E., Cortés MT., y Vélez I. (2004). Seguimiento de egresados de la Facultad de Medicina de la UNAM. Revista de Educación Superior, 7-20.

Navarro Rubén Edel (2003). «Factores asociados al rendimiento académico.»Revista Iberoamericana de Educación.

Belvis Pons Esther, Andrés Moreno Ma. Victoria, y Ferrán Ferrer Julia. (20009). «Los factores explicativos del éxito y fracaso académico en las universidades españolas, en los años del cambio hacia la convergencia Europea.» Revista Española de Educación comparada, no 15, 61-92.

Fayyad U.M., Piatesky-Shapiro G., y Smyth P. (1996) «From Data Mining To Knowledge Discovery: An Overview.» Editado por G. Piatesky-Shapiro, P. Smyth and R. Uthurusamy, U.M. Fayyad. In Knowledge Discovery and Data Mining (AAAI Press/The MIT Press), Menlo Park, CA.

Gatica Lara Florina, Méndez Ramírez Ignacio, Sánchez Mendiola Melchor, y Martínez González Adrían. (2010). «Variables asociadas al éxito académico en los estudiantes de la Licenciatura en Medicina de la UNAM.» Revista de la Facultad de Medicina de la UNAM 53, no 5, 9-11.

Godwa K.C., y Diday E. (March/April 1992). «Simbolic Clustering Using a new Similarity Measure.» IEEE Transaction on Systems, Man and Cybernetic 22, no 2, 368-378.

Gómez Moliné Margarita. (2003). «Algunos factores que influyen en el éxito académico de los estudiantes universitarios en el área de química.» Tesis doctoral, Barcelona.

Zhexue Huang (1998). Extendions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery 2,3, Kluwer Academic Publisher, no 3, 283-304, 1384-5810.

Martínez González A., Urrutia Aguilar M.E., Martínez Franco A.I., Ponce Rosas R., y Gil Miguel A. (2003) «"Perfil del estudiante de posgrado con éxito académico en la UNAM".» Revista de investigación e innovación educativa, no 32, 133-145.

Pressman, Roger S. (2005). Ingeniería del Software: Un Enfoque Práctico. España: McGraw-Hill.

Tinto Vincent. (1992). «El abandono de los estudios superiores: una perspectiva de las causas del abandono y su tratamiento.» Cuadernos de planeación universitaria, México: UNAM (ANUIES) 6, no 2,9-37.

Godwa K.C., y Diday E. (1991). «Symbolic Clustering Using a New Disimilarity Measure.» Pattern Recognition, 567-578.

Hand D.J. (1981),«Discrimination and Classification.» John Wiley & Soon.

Kaufman L., Rousseeuw P. J. (1989), Finding Groups in Data "An Introducction to Cluster Analysis, Wiley series in probability and Mathematical Statistics.

Jain A.J., Dubes R. C. (1988), Algorithms for Clustering Data, Prentice Hall.