Uso de técnicas de clasificación en conglomerados para describir perfiles en grandes bases de datos educativas

Por

Luis Gabriel Jaimes

Tesis sometida en cumplimiento parcial de los requisitos

Para el grado de:

MAESTRÍA EN CIENCIAS EN COMPUTACIÓN CIENTÍFICA

UNIVERSIDAD DE PUERTO RICO

MAYAGÜEZ CAMPUS, 2004

Aprobada por:

| | |
|---|---|
| _____ | _____ |
| Daniel McGee, Ph.D.<br>Presidente, Comité Graduado | Fecha |
| _____ | _____ |
| Julio Quintana, Ph.D.<br>Miembro, del Comité Graduado | Fecha |
| _____ | _____ |
| Keith Wayland, Ph.D.<br>Miembro, del Comité Graduado | Fecha |
| _____ | _____ |
| José Fernando Vega, Ph.D<br>Representante de Estudios Graduados | Fecha |
| _____ | _____ |
| Pedro Vásquez Urbano, Ph.D<br>Chairperson of the Department | Fecha |

# ABSTRACT

This thesis describes the procedures used to identify the natural clusters formed by the students from Calculus that participated in the Quiz Project of the Mathematics Department at the University of Puerto Rico Mayagüez Campus

The principal techniques for clustering and cluster validation are also discussed. In total four, clustering techniques with validation measures were employed

The project was developed in stages. In the first stage data was collected from the Quiz system. For this purpose, an application was developed that inputs files containing quiz results, questionnaires and class results and outputs the information to matrix in which each row represented a student and each column an attribute. During the second stage the information was processed to eliminate unwanted data and weights were assigned to each datum associates with a student. A metric was employed that would help determine the similarities and dissimilarities between students.

To formulate an effective methodology to ascertain the clusters that were present in the actual data, datasets consisting of a number of n-dimensional normal distribution with a variety of means and standard deviations were prepared to simulate real data. The simulated datasets were grouped with the cluster algorithms and validation measures were used to determinate the quality of the grouping. The methodology developed with the simulated data was then applied to find natural cluster existing in the real data.

# RESUMEN

En esta tesis se describe el proceso utilizado para encontrar los conglomerados naturales en los que se encuentran agrupados los estudiantes de Cálculo I que participan en el proyecto QUIZ del Departamento de Matemáticas de la UPRM. Además, se muestra un panorama general de las principales técnicas de agrupación y validación de conglomerados. Para este propósito se utilizaron cuatro de los algoritmos más representativos de estas técnicas y cuatro medidas de validación. El proyecto se desarrolló en varias etapas. En la primera se extrajo la información del sistema Quiz, para lo cual se creó una aplicación que extrae información de archivos de texto en diferentes formatos y con estos datos crea una matriz donde cada fila corresponde a un estudiante y cada columna a sus atributos. En la siguiente etapa se procesó la información mediante la eliminación de datos y la asignación de pesos. Luego se adoptó una métrica para determinar lo similar o disimilar que son los estudiantes entre sí. Como se desconocen las clases a las cuales pertenecen los alumnos y el número de ellas, se diseñó una metodología para identificarlas. Esta metodología consiste de la aplicación de algoritmos de agrupamiento a datos generados aleatoriamente normalmente distribuidos que pretenden simular los datos reales. Posteriormente se midió la calidad de los conglomerados formados con las medidas de validación y se compararon estos conglomerados con los conglomerados reales. De esta forma se pudo establecer el comportamiento de los algoritmos con diferentes estructuras de datos y el grado de confiabilidad de las diferentes medidas de validación. Finalmente se aplicó esta metodología para encontrar los conglomerados naturales subyacentes en los datos reales.

## CONSENTIMIENTO

Por este medio autorizo a la Biblioteca del Recinto Universitario de Mayagüez a permitir copiar en parte o en su totalidad al documento de trabajo realizado para la obtención de mi grado. La copia parcial o total del mismo será únicamente para propósitos de INVESTIGACIÓN.

## CONSENT

I do herby authorize the Library of University of Puerto Rico at Mayagüez to allow partial or complete copying of this document for research purposes.

# RECONOCIMIENTOS

A Dios que me acompaña cada día con su amor y misericordia.

A mi consejero el Dr. Daniel McGee por su apoyo constante y valiosos consejos durante la preparación y ejecución del presente trabajo.

A la Universidad Puerto Rico Recinto Universitario de Mayagüez por su apoyo económico durante el transcurso de la maestría.

# CONTENIDO

**LISTA DE TABLAS**

# LISTA DE FIGURAS

# CAPÍTULO 1: INTRODUCCIÓN

Para apoyar el aprendizaje en el área de Precálculo y Cálculo, el Departamento de Matemáticas de la UPRM creó el proyecto QUIZ. Este sistema esta orientado al mejoramiento del nivel académico de los estudiantes, mediante la evaluación en línea y apoyo al profesor en el aula de clase. El proyecto tiene el propósito de no solamente mejorar la educación sino también de crear datos que puedan ser utilizados para evaluar el mejoramiento de la enseñanza de estas asignaturas. A continuación se hará una descripción del sistema que nos permitirá conocer el objeto de nuestro estudio.

## 1.1 Panorama general del sistema que provee los datos para el análisis de conglomerados.

Las cinco principales componentes del sistema QUIZ son:

(i) Un laboratorio donde los estudiantes toman las pruebas cortas.

(ii) El *Internet Quiz Generating System* que es la herramienta que se utiliza para administrar las pruebas cortas.

(iii) Los tutoriales a los cuales son referidos los estudiantes.

(iv) Bases de datos de preguntas que son usadas para las pruebas cortas reales y de práctica.

(v) Cuestionarios que reflejan las actitudes y opiniones de los estudiantes.

### 1.1.1 Laboratorio

Las pruebas cortas de práctica pueden ser tomadas en un ambiente no controlado tal como el laboratorio, en la universidad o en la casa. Cuando los estudiantes toman pruebas cortas que cuentan para su nota (reales) deben hacerlo en un ambiente controlado, es decir, el laboratorio, donde su identidad pueda ser verificada.

### 1.1.2 Bases de datos de preguntas.

Para administrar las pruebas reales y prácticas, se han desarrollado bases de datos de preguntas grandes y detalladas para cubrir muchos de los temas vistos en el salón de clase, y repaso de temas de Precálculo y Cálculo. Estas preguntas pueden ser de escogencia múltiple; pueden requerir una respuesta numérica; que el estudiante introduzca una fórmula o una gráfica.

Los siguientes son los criterios utilizados en el desarrollo de estas bases de preguntas.

- Aproximadamente de 9 a 12 bases existirán para cada tema. La organización está dada por *Enfoque del tópico – Grado de dificultad*. Por ejemplo, el tema "integrales geométricas fáciles", contendrá preguntas concernientes a integrales que tienen un enfoque geométrico y son fáciles.

- Cada tema debe contener bases de datos con cuatro enfoques: algebraico, geométrico, numérico y aplicado. Si un profesor no desea que todos los enfoques estén en una prueba corta, no necesita seleccionar todas las bases de datos de un tema.
- Los profesores podrán editar, agregar o remover el contenido de las bases de datos como mejor se ajuste a sus necesidades.
- Los profesores pueden contribuir a la base de datos de preguntas central.
- Todas las bases de datos en este banco central de preguntas están disponible sin costo alguno para todas las instituciones interesadas.

### 1.1.3 Software

El *Quiz Generating System* que usa Internet como herramienta trabaja de la siguiente forma:

(1) El profesor llena una forma electrónica indicando el contenido de la prueba corta; (2) Los estudiantes reciben una prueba corta de práctica que ha sido aleatoriamente generada de la base de preguntas y es única para cada estudiante; (3) Los estudiantes responden a la prueba corta y ésta es sometida con sus respuestas; (4) El computador corrige la prueba y genera un informe para el estudiante. El informe contiene la siguiente información: (i) si aprobó; (ii) enlaces electrónicos donde ellos pueden repasar los temas donde fallaron y (iii) las respuestas incorrectas. Si aprueban la prueba, serán promovidos para tomar la prueba real que será administrada en el laboratorio, de manera supervisada.

Este “software” no tiene costo para otras instituciones. Es fácil de usar con múltiples secciones, cursos y exámenes departamentales escritos. En Mayagüez se esta aplicando actualmente en cursos que totalizan aproximadamente mil alumnos pertenecientes a las secciones de Precálculo I, Precálculo II, Cálculo I, Cálculo II y Cálculo III. El coordinador del curso diseña la prueba semanal y los profesores envían informes semanales a sus estudiantes, y un informe final al culminar el semestre. Esto no requiere ningún esfuerzo de parte de los profesores excepto incorporar el porcentaje asignado de las pruebas de Internet en los notas finales.

### 1.1.4 Tutoriales

Después de que los estudiantes toman la prueba de práctica, el sistema determina el tema que necesitan reforzar y los envía a una referencia apropiada, que generalmente es un tutorial electrónico. Sin embargo, el profesor puede sustituirlo con cualquier otro enlace o con la referencia que estime conveniente. La Universidad de Puerto Rico ha estádo usando cálculo visual para la creación de tutoriales.

### 1.1.5 Cuestionarios

Con los cuestionarios se pretende conocer las actitudes y opiniones de los estudiantes con respecto al sistema. Además se indaga sobre hábitos de estudio, tal como número de horas diarias o semanales que dedican al estudio de la clase, expectativas de

recibir cierta nota, grado de claridad de las preguntas usadas en las pruebas, grado de dificultad que encuentran en las preguntas de las pruebas, etc. Esto permite conocer los hábitos de estudio de los estudiantes y el tiempo dedicado a la clase. Además permite mejorar el sistema continuamente mediante el ajuste de las bases de datos, para determinar si existe o no relación con los temas expuestos en clase. Generalmente se aplican dos cuestionarios durante el semestre, cada uno con aproximadamente 12 preguntas, de las cuales 10 son de interés para esta tesis.

Creemos que los datos que arroja el sistema contienen suficiente información para realizar un estudio preliminar encaminado a distinguir entre los diferentes perfiles en que se agrupan los estudiantes de Cálculo I de la UPRM.

## 1.2 Objetivo

Desarrollar un análisis preliminar y los algoritmos necesarios para determinar los conglomerados naturales en que se encuentran agrupados los estudiantes que toman el curso de Cálculo I que se ofrece en el RUM y describir los perfiles asociados a estos conglomerados.

## 1.3 Bosquejo general de la tesis

Se han realizado algunos estudios para distinguir perfiles de estudiantes [1] y todos ellos utilizan el enfoque de clasificación supervisado. En esta tesis se pensó que no se

debería suponer nada acerca de los grupos a los cuales pertenecen los estudiantes y que sería prejuiciado intentar hacerlo, pues un criterio personal sesgaría los resultados impidiendo ver cosas subyacentes en los datos. Por esta razón se escogió el enfoque de clasificación no supervisado y específicamente técnicas de agrupación en conglomerados. Pero antes de poder utilizarlos, estos datos deben obtenerse del sistema y extraer de ellos las porciones de información necesarias. Para esta etapa del proyecto se construyó una aplicación que toma como entrada todas la fuentes de datos discutidas en la secciones anteriores (*pruebas prácticas en Internet, pruebas reales en Internet, trabajo en clase y respuestas a cuestionarios*) y su salida es una matriz de datos, en donde cada estudiante es una fila y cada uno de sus atributos es una columna. Como es de suponer, no todos los estudiantes tienen sus atributos completos, pues muchos de ellos se dieron de baja en el transcurso del semestre, o dejaron de tomar algunas de las pruebas reales, o alguno de los cuestionarios, lo que implica la existencia de muchos datos faltantes. Por esta razón se eliminó de la matriz de récords a los estudiantes que se dieron de baja, y además se eliminaron los récords de todos los estudiantes que no tuvieron completos sus atributos. Por otra parte todas las variables no tienen la misma importancia, así que *asignamos ponderaciones* a cada conjunto de atributos. Por ejemplo, no puede tener el mismo peso una prueba de práctica en Internet que el examen final tomado en el aula de clase.

Como utilizamos el enfoque no supervisado, desconocemos el número de grupos y los grupos a los que pertenecen los estudiantes. Éste es un problema abierto en análisis de conglomerados pues en problemas reales, la mayoría de las veces se desconocen los grupos a los que pertenece cada observación. Por esta razón se decidió diseñar una

metodología que permitiera determinar la proporción entre la dispersión de los datos alrededor de los centros y la distancia entre los centros necesaria para determinar cuando un conglomerado es claro. Con este propósito se escogieron cuatro algoritmos para agrupar en conglomerados, dos de ellos de particionamiento, K-means y PAM y dos jerárquicos, uno jerárquico aglomerativo AGNES y otro jerárquico divisivo DIANA. Estos algoritmos son representativos de las principales clases de algoritmos para agrupar en conglomerados. Cada uno de ellos se describe en detalle en el Capítulo 5. Además se usaron cuatro medidas para validar la calidad de los conglomerados; éstas fueron: el *Average Silhouette*, el *Mean Split Silhouette*, el coeficiente de *Calinski-Harabasz* y el coeficiente *Davies-Bouldin*. Cada una de ellas se describe en detalle en el Capítulo 6. Se realizaron simulaciones con datos generados siguiendo una distribución normal; estos datos así generados tratan de simular los datos reales. Luego se agruparon los datos a través de los cuatro algoritmos anteriormente mencionados y se midió la calidad de los conglomerados utilizando las cuatro medidas de validación. Este procedimiento da una idea del comportamiento de los algoritmos con las diferentes estructuras de datos, y además permite determinar la confiabilidad de las medidas de validación. Se escogió la métrica euclidiana para medir la distancia dentro y entre los conglomerados. Una vez hechas las simulaciones se procede a agrupar los datos reales. Para ello se busca consistencia entre todos los métodos de agrupación y se escogen los conglomerados que se manifiesten simultáneamente en agrupaciones hechas por todos los métodos, teniendo en cuenta las relaciones de densidad de datos alrededor de los centros y distancia entre centros encontradas en las simulaciones.

# CAPÍTULO 2: TRABAJOS PREVIOS EN CLASIFICACIÓN CON DATOS EDUCATIVOS

Tradicionalmente el proceso de clasificación se ha hecho sabiendo de antemano los grupos a los que deben pertenecer los alumnos (clasificación supervisada). En este proyecto se quiere que los datos se clasifiquen a sí mismos, es decir, no se presume nada acerca de cómo se deberían clasificar (clasificación no supervisada). Esta técnica de clasificación es novedosa para este campo en particular (determinación de perfiles educativos), pues aunque se usa habitualmente en campos como microarrays [2], reconocimiento de patrones [3] y tratamiento de imágenes, entre otros, no se encontró referencias de que se haya usado para determinar perfiles educativos.

Es común clasificar a los estudiantes dependiendo de algún criterio, por ejemplo, *el tiempo que ha transcurrido desde su iniciación*. Este criterio los puede asignar a una de las siguientes clases: Freshman, Sophomore, Junior, Senior. Los puntos que ha obtenido al presentar un examen, los puede clasificar dentro de A, B, C, D o F, etc. También es común encontrar estudios que utilizan clasificación supervisada para tratar de predecir los resultados finales de los estudiantes, con base en bases de datos de resultados anteriores.

Uno de los problemas que se aborda en clasificación educativa es tratar de predecir las notas finales de un estudiante. El siguiente estudio fue realizado en este sentido utilizando un enfoque supervisado: *Genetic Algorithms for Data Mining Optimization in an Educational Web-based System* [4]. En este estudio se pretende predecir las notas finales de los estudiantes con base en las características extraídas de

grandes bancos de datos de un sistema educativo que utiliza un sistema Web. En este caso se usó una combinación de clasificadores, entre ellos el llamado *Clasificador Cuadrático Bayesiano* (Quadratic Bayesian Classifer), *El Primer vecino más cercano* (1-Nearest Neighbord) y el *K-ésimo vecino más cercano* (K-Nearest Neighbor). Las variables en este trabajo son similares a las de esta tesis, y además también provienen de un sistema Web. Sin embargo se utiliza un enfoque de clasificación supervisada para lograr el objetivo.

Se ha revisado cuidadosamente la situación actual y no se han encontrado trabajos anteriores en donde se busquen perfiles educativos utilizando técnicas de agrupación en conglomerados, por lo que este trabajo es innovador en este sentido.

# CAPÍTULO 3: METODOLOGÍA

En este proyecto se consideran varias etapas: La primera etapa corresponde a la extracción y manejo de datos; la segunda etapa consiste en el desarrollo de una metodología que permita encontrar una relación entre la dispersión de los datos alrededor de los centros y la distancia entre centros necesaria para identificar correctamente los conglomerados; y la tercera etapa consiste en la utilización de esta relación para encontrar los conglomerados que son consistentes a través de todas las agrupaciones hechas con los diferentes métodos sobre los datos reales. Se determina entonces cuáles están claramente agrupados y cuáles no. Luego se describe un perfil de cada conglomerado.

## 3.1 Primera Etapa: Manejo de datos

En la primera etapa, se realizó el procesamiento de los datos. Los datos que se utilizaron en este proyecto son los resultados de las pruebas que utilizando Internet toman los estudiantes de Cálculo I del departamento de matemáticas del RUM. Además, se usaron los resultados de cuestionarios y encuestas a los que fueron sometidos los estudiantes para determinar las aptitudes y hábitos al responder a las pruebas y a los exámenes en clase. Estos datos provienen de diversas fuentes, algunos de ellos procesados por CGI's (Common Gatway Interface) que están en forma de grandes

archivos de texto; los cuales contienen diversos datos, muchos de los cuales no son de interés para este estudio.

Para extraer de diversos archivos la información de nuestro interés, se construyó una aplicación orientada a objetos en C++. Ésta toma como entrada grandes archivos de texto, los procesa y luego entrega como salida una matriz de datos, donde cada fila u observación representa el récord de un estudiante, y cada columna representa una característica o variable. En este sentido se considera el récord de cada estudiante como un punto en el espacio n-dimensional donde cada dimensión corresponde a una característica de dicho estudiante.

## 3.2 Segunda etapa: Determinación de la relación dispersión alrededor de los centros y distancia entre centros

Cuando se hace clasificación no supervisada y para ello se utilizan conglomerados (clusters), existe el problema abierto de confiar en los conglomerados identificados, pues dependiendo de la naturaleza de las observaciones (en este caso con alta dimensionalidad), el tipo de algoritmo y la métrica que se utilice, los resultados pueden ser significativamente diferentes. Esto hace que la labor de clasificar no sea fácil y muchos investigadores opten por escoger otro tipo de clasificación. Por esta razón se decidió elaborar una metodología que permitiera determinar todos los conglomerados naturales subyacentes en los datos. Con este propósito se utilizaron simulaciones. Inicialmente se crearon datos normalmente distribuidos, con diferentes desviaciones estándar que van desde un mínimo, donde las medidas de validación nos muestran que

todos los algoritmos están agrupando en los grupos correctos (datos densamente concentrados alrededor de los centros), hasta el punto donde todos ellos se pierden, es decir, hasta que las medidas de validación muestran que ningún algoritmo está agrupando de manera correcta. Además, inicialmente se varió la dimensión del conjunto de observaciones de 4 a 20 variables. Esto da una idea del comportamiento de los algoritmos con diferentes estructuras de datos.

Para agrupar los datos en conglomerados, se utilizaron cuatro algoritmos: k-means, PAM (Partitioning Around Medoids), AGNES (Agglomerative Nesting) y DIANA (Divisive Analysis). Los dos primeros son métodos de particionamiento, el tercero es jerárquico aglomerativo y el cuarto, jerárquico divisivo. Para estimar el número de conglomerados que se encuentran naturalmente en los datos se utilizaron cuatro medidas: *Average Silhouette*, *Mean Split Silhouette,* el coeficiente de *db* y el coeficiente *de Calinski.* La siguiente metodología fue utilizada en las simulaciones: se realizaron 25 simulaciones por terna de datos es decir, (*número de cluster*, *desviación estándar*, *dimensión de los datos* (número de variables)), manteniendo fija una de ellas hasta agotar las demás. Un ejemplo de una de las 25 simulaciones por terna se muestra en la Tabla 1 y las siglas de los métodos están resumidas en la Tabla 2.

**Tabla 1**. Primera etapa del proceso de simulación

| Sd = 0.5 | | | numvar=4 | | | numclust=3 | |
|---|---|---|---|---|---|---|---|
| No. | Ps | As | Ds | ks | Pc | Pd | MS |
| 2.0000 | 0.5347 | 0.5918 | 0.5918 | 0.5918 | 126.1889 | 1.1689 | 0.6535 |
| 3.0000 | 0.6526 | 0.6526 | 0.6526 | 0.6526 | 402.0196 | 0.5629 | 0.1900 |
| 4.0000 | 0.4988 | 0.5444 | 0.5036 | 0.5031 | 296.0096 | 1.4533 | 0.2464 |
| 5.0000 | 0.3359 | 0.3647 | 0.3954 | 0.3689 | 239.8968 | 1.8392 | 0.3028 |
| 6.0000 | 0.1748 | 0.3561 | 0.2631 | 0.3637 | 208.4636 | 1.9171 | 0.2381 |
| 7.0000 | 0.1992 | 0.3532 | 0.2623 | 0.3457 | 186.1865 | 1.5862 | NA |

**Tabla 2**. Siglas utilizadas en las simulaciones

| ***Sigla*** | ***Conglomerados*** | ***Medida*** |
|---|---|---|
| ps | PAM | Silhouette |
| as | AGNES | Silhouette |
| ds | DIANE | Silhouette |
| ks | k-Means | Silhouette |
| pc | PAM | Calinski |
| pd | PAM | DB |
| ms | PAM | Mean Split Silhouette |

En este ejemplo *ps* (PAM/ Silhouette) muestra que para particiones de 2 a 7 la mejor es 3, pues maximiza la función objetivo, es decir, en 3 conglomerados está el mayor valor de la función. Los datos de la Tabla 1 quedan resumidos en el primer renglón de la tabla 3 que a su vez agrupa 25 simulaciones como la anterior.

**Tabla 3**. Índices de acierto

| sd = 0.5 Numvar=4 cluster=3 corrida=1 | | | | | | |
|---|---|---|---|---|---|---|
| ps | as | ds | ks | pc | pd | ms |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 4 |
| 3 | 3 | 3 | 2 | 3 | 3 | 4 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 4 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 4 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 2 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 6 |
| 3 | 3 | 3 | 2 | 3 | 3 | 4 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 2 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 4 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 2 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 6 |
| 3 | 3 | 3 | 3 | 3 | 3 | 5 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 3 | 3 | 3 | 3 | 3 | 3 | 3 |

En la tabla 3 se observa el número de conglomerados que arroja cada indicador de *agrupación/validación*, que ha sido aplicado a un conjunto de datos creado con cuatro variables y tres conglomerados. Las medias de este conjunto están separadas a dos unidades la una de la otra y los datos están normalmente distribuidos con una desviación estándar de 0.5. En este ejemplo se puede observar que los indicadores *ps* y *as* mostraron el número correcto de conglomerados en las 25 simulaciones.

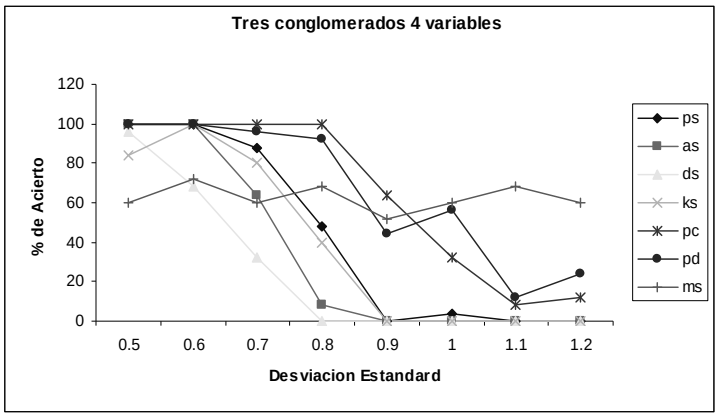


**Figura 1**: Ejemplo del proceso de simulación

La Figura 1 muestra gráficamente la tabla anterior en términos de porcentajes. Aquí la Tabla 3 representa sólo los puntos que corresponden a una desviación estándar de 0.5. Los detalles de las simulaciones se explican en el Capítulo 7.

Los criterios fundamentales que determinan la capacidad de identificar un conglomerado son la densidad de los datos alrededor de los centros y la distancia entre centros. Si los datos son densos alrededor de los centros y la distancia entre los centros es grande, es fácil identificar los conglomerados. Pero si los datos no son densos alrededor de los centros y los centros son cercanos, es difícil identificarlos. Lo que se busca con está metodología es encontrar la densidad relativa a la distancia entre centros necesaria para identificar conglomerados correctamente.

## 3.3 Tercera etapa: Identificación de los conglomerados en los datos reales y descripción de perfiles

En está etapa buscamos identificar los conglomerados que se manifiestan en los datos reales y describir los perfiles en términos de sus variables. Para este propósito se busca consistencia entre todos los métodos, es decir se buscan los conglomerados que se manifiestan simultáneamente en todas las agrupaciones hechas por los diferentes métodos de clasificación. Por ejemplo, si se agrupa en *n* conglomerados, utilizando un método, se busca cuántos de esos *n* conglomerados se manifiestan en agrupaciones hechas por los demás métodos. Por ejemplo, se busca en cuáles agrupaciones se manifiesta el conglomerado 1. Podría ser que coincida con el conglomerado 3 de una agrupación hecha con PAM y al conglomerado 5 de una agrupación hecha con KMEANS y al conglomerado 7 de una agrupación hecha con AGNES y al conglomerado 8 de una agrupación hecha con DIANA. Para este propósito será muy útil conocer la relación entre dispersión alrededor de los centros y la distancia entre centros necesaria para identificar correctamente los conglomerados.

Una vez determinados los conglomerados consistentes con todos los métodos se dará un perfil en términos de las variables de cada conglomerado.

# CAPÍTULO 4: MANEJO DE DATOS

Una de las etapas más importantes es el procesamiento de los datos. Este consiste inicialmente en la extracción de la matriz de observaciones que será usada como entrada de los algoritmos de clasificación.

Los datos que arroja el sistema “QUIZ” son grandes archivos con información semanal de los estudiantes de los cursos antes mencionados. Estos archivos contienen mucha información que no es de interés en este estudio, así que deben ser depurados y debe extraerse la información necesaria para completar los datos de un estudiante. Por ejemplo, para completar la fila que dentro de una matriz representa un estudiante, se necesita recolectar la información de los resultados de las pruebas de práctica, y el numero de estás varía, pues sólo puede tomar una prueba real hasta que apruebe una de las pruebas de practica. Además se necesita recolectar semanalmente los resultados de las pruebas reales por Internet. Este proceso se repite de 9 a 12 veces dependiendo del semestre. También se recolectan las respuestas de los cuestionarios una o dos veces por semestre y los resultados de los exámenes parciales y final hechos en el aula de clase. Esto nos da una idea de la complejidad del manejo de los datos y de la naturaleza de la aplicación que debió construirse para manejarlos.

## 4.1 La aplicación

Para construir está aplicación se escogió el enfoque orientado a objetos, pues este es el que representa más fielmente la naturaleza de los datos. Además permite añadir y remover fácilmente nuevos módulos y clases sin afectar el funcionamiento de la aplicación, por ejemplo, permite agregar cursos, remover semanas, agregar o remover cuestionarios, etc.

### 4.1.1 Manejo y administración de datos

Para el acceso y organización de este gran volumen de información se utilizó el modelo orientado a objetos. Se utilizá está técnica de programación pues aumenta la velocidad de desarrollo y facilita su mantenimiento [5]. Se ha creado una estructura de árbol para manejar la información. Este esquema para extraer la información brinda tres características importantes: rapidez, fácil almacenamiento y una forma sencilla de realizar cambios.

*a. Rapidez:* La información se toma de archivos de texto producto de los CGI's que capturan información de los estudiantes al momento de presentar las pruebas. Está información se extrae utilizando una aplicación construida en C++. El tipo de estructura de datos, el paradigma y el lenguaje de programación permiten que el proceso sea rápido y eficiente, dada la complejidad de las relaciones que existen entre los objetos (estudiante, curso, sección, semana, prueba de práctica y real, cuestionario, examen parcial y final).

*b. Almacenamiento*: La información se procesa y almacena en archivos planos. Estos archivos contienen líneas con las variables que se han capturado al momento de presentar una prueba. El registro de cada estudiante es una línea representada en términos de variables, es decir representa un punto de n-dimensionalidad.

*c. Manejo de Cambios:* Tal vez esto es uno de los requerimientos más importantes, dada la naturaleza dinámica del proceso. Por está razón se escogió una estructura de árbol pues es fácil podar y adherir subárboles, proceso que corresponde a las diferentes semanas de clases, la adición y remoción de cursos, etc.

### 4.1.2 Estructura de los datos

La Figura 2 ilustra la forma como se estructuraron los datos.

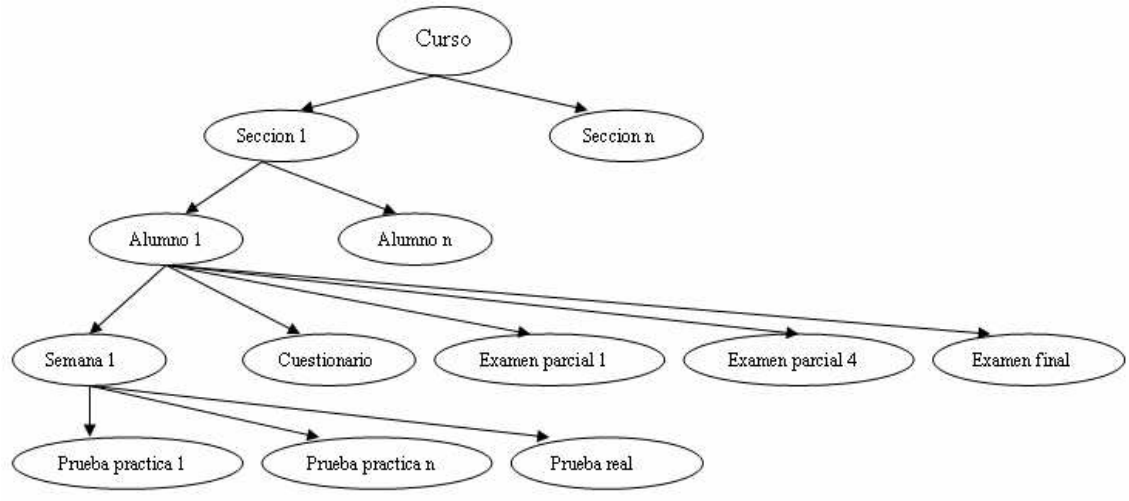


**Figura 2**. Estructura de los datos

La Figura 3 ilustra el proceso de procesamiento de datos.

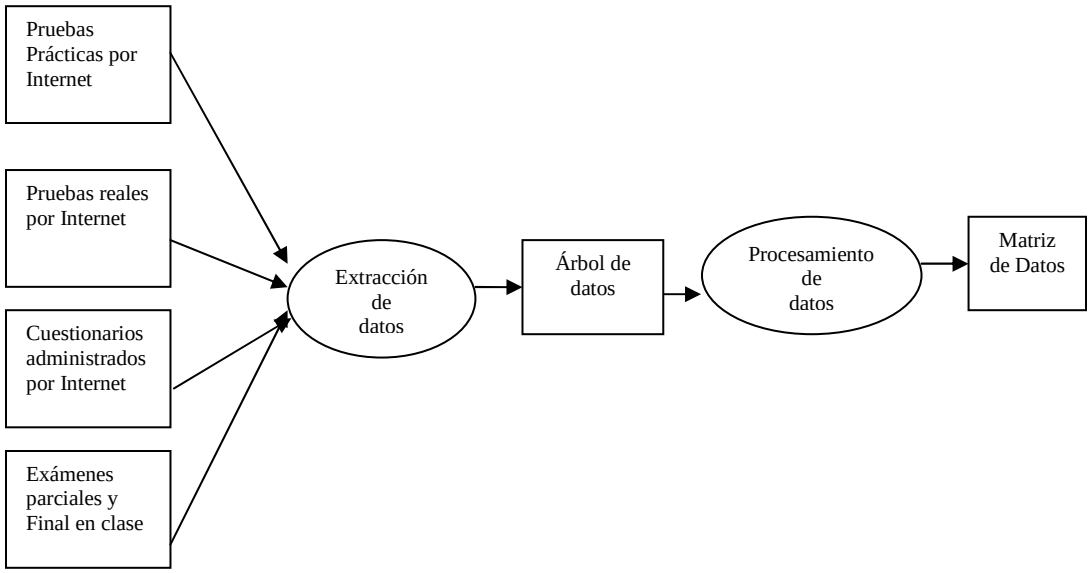


**Figura 3**. Flujo del procesamiento de datos

## 4.2 Eliminación de datos

Una vez que la aplicación ha extraídos del sistema los datos necesarios, se tiene una matriz de $n \times p$, pero no todos los estudiantes tiene sus atributos completos; muchos de ellos se han dado de baja durante el transcurso del semestre y esto hace que falten muchos atributos. Por otra parte si un estudiante no ha tomado los cuestionarios, o ha dejado de tomar alguna de las pruebas cortas reales o de práctica, se decidió eliminarlo de la matriz, pues cualquiera de estás situaciones supone al menos 8 datos faltantes (Hay por lo menos 10 preguntas de interés en los cuestionarios y cada examen de práctica tiene por lo menos 8 preguntas).

## 4.3 Asignación de ponderaciones

Como todos los datos no tienen la misma importancia, se asignó una valoración diferente a cada conjunto de ellos. Por ejemplo, se consideró que las pruebas presentadas en el aula de clase son más importantes que las pruebas de práctica presentadas en Internet y por está razón se les asignó una ponderación superior. Las asignaciones de peso se hicieron de la siguiente manera: a cada una de las preguntas de las pruebas prácticas y reales en Internet se les asignó un peso de 1; a cada una de las preguntas de los cuestionarios tomados en Internet un peso de 2; a las cuatro pruebas reales tomadas en el aula de clase, un peso de 50, y al examen final también un peso de 50. El criterio para la asignación fue subjetivo.

# CAPÍTULO 5: PANORAMA GENERAL DE LAS TÉCNICAS PARA AGRUPAR EN CONGLOMERADOS.

Al igual que el análisis factorial, el análisis de conglomerados estudia todo un conjunto de relaciones interdependientes. Este análisis no hace ninguna distinción entre variables dependientes y variables independientes. En vez de ello, se calculan las relaciones interdependientes de todo el conjunto de variables [6].

## 5.1 ¿En qué consisten los métodos de agrupación en conglomerados?

Estos métodos de agrupamiento caen en el campo de la clasificación no supervisada y se utilizan para clasificar los objetos o casos en grupos relativamente homogéneos llamados conglomerados o “clusters”. Los objetos en cada grupo tienden a ser similares entre sí y diferentes a los objetos en otros grupos. Este análisis se conoce también como análisis de clasificación o taxonomía numérica.

## 5.2 Pasos en el análisis de conglomerados

La siguiente gráfica ilustra el proceso de análisis de conglomerados que se utiliza en este proyecto.

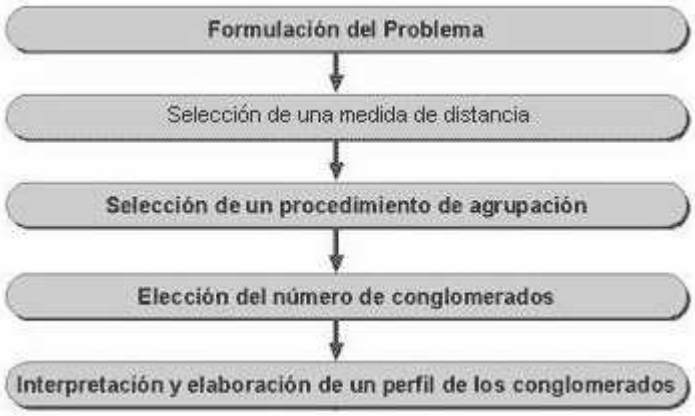


**Figura 4.** Pasos en el proceso de análisis de conglomerados

### 5.2.1 Formulación del Problema

Quizá la parte más importante de la formulación del problema es la selección de las variables en las que se basa la agrupación. La inclusión de una o más variables irrelevantes puede distorsionar una solución de agrupación que de otra forma podría ser útil. En este caso las variables son los atributos del estudiante (pruebas prácticas por Internet, pruebas reales por Internet, cuestionarios por Internet y exámenes en el aula de clase). De estos atributos ya se ha hablado en secciones anteriores.

### 5.2.2 Selección de la Medida de Distancia

Ya que el objetivo de estás técnicas es agrupar objetos similares, se necesita alguna medida para evaluar las diferencias y similitudes entre objetos. La estrategia más común consiste en medir las similitudes en términos de la distancia entre los pares de objetos. Los objetos con distancias reducidas entre ellos son más parecidos entre sí que

aquellos que tienen distancias mayores [7]. Existen varias formas de calcular las distancias entre dos objetos. La medida de similitud que se utiliza con mayor frecuencia es la distancia euclidiana

$$D_2 = \sqrt{(\sum_{i=1}^{M}(x_i - y_i)^2)} \quad (5.1)$$

o su cuadrado que es un caso particular de la distancia *Minkowski* o en norma $L_p$

$$D_p(\mathbf{x},\mathbf{y}) = (\sum_{i=1}^{M}(x_i - y_i)^p)^{1/p} \quad (5.2)$$

También están disponibles otras medidas de distancia. La distancia Manhattan o de Calles Urbanas entre dos objetos que es la suma de las diferencias absolutas en los valores para cada variable. Está dada por:

$$D_1 = \sum_{i=1}^{M} | x_i - y_i | \quad (5.3)$$

La distancia de Chebychev entre dos objetos es la diferencia absoluta máxima en los valores para cualquier variable

$$D_\infty = \max_{1 \le i \le M} | x_i - y_i | \quad (5.4)$$

Dada Q = $(Q_{ij})$ una matriz cuadrada *MxM* definida positiva de pesos, entonces la distancia cuadrática entre **x** y **y** está dada por:

$$D_Q(\mathbf{x},\mathbf{y}) = [(\mathbf{x}-\mathbf{y})'Q(\mathbf{x}-\mathbf{y})]^{1/2} = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{M}(x_i - y_i)Q_{ij}(x_j - y_j)} \quad (5.5)$$

Un caso particular de está distancia es cuando $Q = V^{-1}$, siendo V la matriz de covarianza entre **x** y **y**. En este caso la distancia es conocida con el nombre de distancia Mahalanobis. En este proyecto se uso la distancia euclidiana pues es la adecuada para mantener la compatibilidad con el futuro desarrollo de un modulo de visualización. Sin embargo se considera que en proyectos futuros deberían hacerse simulaciones, para buscar la métrica que arroje mejores resultados en términos de identificación de conglomerados y que más favorezca la asignación de pesos.

### 5.2.3 Selección del Procedimiento de Aglomeración

Se escogieron cuatro algoritmos de agrupación, dos no jerárquicos; *K-means* y *PAM,* y dos jerárquicos: uno jerárquico aglomerativo, *AGNES,* y otro jerárquico divisivo, *DIANA.*

### 5.2.4 Clasificación de los métodos de conglomerados

Las figura 5 ilustra una clasificación de los métodos aglomerativos según [8].

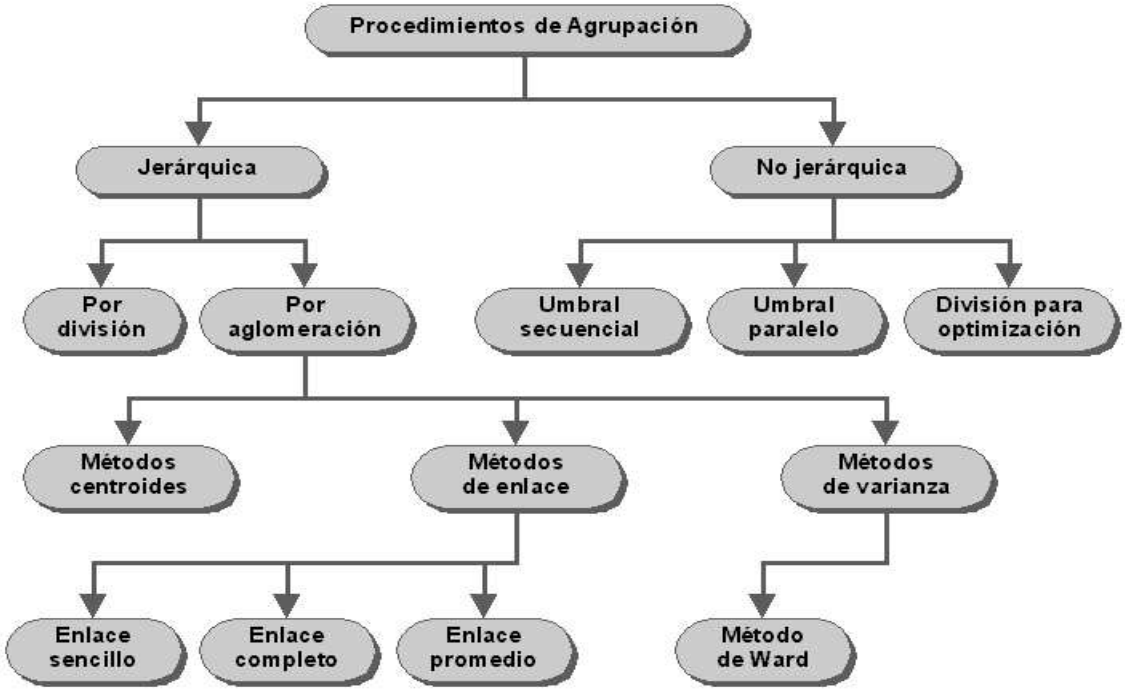


**Figura 5.** Clasificación de los métodos de agrupación

### 5.2.4.1 Métodos no jerárquicos (particionamiento)

El conjunto de datos es particionado en un número pre-especificado de conglomerados K, y luego iterativamente se asignan las observaciones a los conglomerados hasta se satisface algún criterio de parada (función a optimizar), por ejemplo, que la suma de cuadrados dentro de los conglomerados sea mínima.

Entre los algoritmos de este tipo utilizados en este proyecto se encuentran:

i) K-means; y

ii) Particionamiento alrededor de los medoides (PAM, Kaufman and Rosseeuw 1990)

### 5.2.4.1.1 K-Means

Visto como un problema de optimización, el objetivo del algoritmo k-means es minimizar la distancia de los elementos dentro de cada conglomerado, al mismo tiempo que maximiza la distancia de los elementos que caen en diferentes conglomerados.

El algoritmo es como sigue:

*Input*: Un conjunto de datos S y k número de conglomerados a formar;
*Output*: L una lista de los conglomerados en que caen las observaciones de S

*Paso 1*. Seleccionar los centroides iniciales de los K conglomerados: $c_1$, $c_2$, $c_3$, ..., $c_K$.

*Paso 2*: Asignar cada observación $x_i$ de S al conglomerado **C(i)** cuyo centroide c(i) está más cerca a $x_i$. Es decir,

$$C(i) = \arg\min_{1 \le k \le K} \|x_i - c_k\| \quad (5.5)$$

*Paso 3*. Para cada uno de los conglomerados se recalcula su centroide con base en los elementos que lo conforman y se minimiza la suma de cuadrados dentro del conglomerado. Es decir,

$$WSS = \sum_{k=1}^{K} \sum_{C(i)=k} \| \mathbf{x}_i - \mathbf{c}_k \|^2 \quad (5.6)$$

*Paso 4*. Ir al paso 2 hasta que se consiga convergencia, es decir, hasta que la asignación de observaciones en los grupos no cambie con respecto a la iteración anterior.

Los k centroides iniciales pueden ser considerados de varias maneras

a) Usando las primeras k observaciones,
b) Eligiendo aleatoriamente k observaciones,
c) Tomando cualquier partición al azar en k conglomerados y calculando sus centroides.

El algoritmo k-means es de cálculo rápido y puede trabajar bien con datos faltantes (missing values), pero es sensible a datos atípicos (outliers) [9].

En este proyecto hemos dejando que los centros sean elegidos al azar. El siguiente es un ejemplo sencillo en una dimensión que ilustra en términos generales como funciona el método.

*Ejemplo:* en la primera columna se encuentra la posición del elemento y en la segunda su valor. Se han elegido inicialmente 2 centroides, ubicados en las posiciones 2 y 7. En la columna con etiqueta *dist 1* se ha registrado la distancia de cada objeto al primer centroide. De igual forma, en la siguiente columna se ha registrado la distancia de cada objeto al siguiente centroide. Luego se han escogido las distancias mínimas, y en la ultima columna de la tabla 3 ser realiza la asignación de elementos a cada uno de los conglomerados.

**Tabla3**. Ejemplo de K-Means

| ***Número*** | ***objeto*** | ***dist 1*** | ***dist 2*** | ***mínima d.*** | ***cluster*** |
|---|---|---|---|---|---|
| 1 | 9 | 1 | 2 | 1 | 1 |
| ***2*** | ***10*** | 0 | 3 | 0 | 1 |
| 3 | 4 | 6 | 3 | 3 | 2 |
| 4 | 5 | 5 | 2 | 2 | 2 |
| 5 | 9 | 1 | 2 | 1 | 1 |
| 6 | 3 | 7 | 4 | 4 | 2 |
| **7** | **7** | 3 | 0 | 0 | 2 |
| *8* | *25* | 15 | 18 | 15 | 1 |
| 9 | 8 | 2 | 1 | 1 | 2 |
| 10 | 0 | 10 | 7 | 7 | 2 |

Se recalculan los centros, como el promedio de las distancias dentro de cada conglomerado. Los nuevos centroides son: **4.25, 2.83**

**Tabla 5.** Nueva iteración de K-Means

| ***Número*** | ***objeto*** | ***dist 1*** | ***dist 2*** | ***mínima d.*** | ***cluster*** |
|---|---|---|---|---|---|
| 1 | 9 | 4.75 | 6.17 | 4.75 | 1 |
| ***2*** | ***10*** | 5.75 | 7.17 | 5.75 | 1 |
| 3 | 4 | 0.25 | 1.17 | 0.25 | 1 |
| 4 | 5 | 0.75 | 2.17 | 0.75 | 1 |
| 5 | 9 | 4.75 | 6.17 | 4.75 | 1 |
| 6 | 3 | 1.25 | 0.17 | 0.17 | 2 |
| ***7*** | ***7*** | 2.75 | 4.17 | 2.75 | 1 |
| *8* | *25* | 20.75 | 22.17 | 20.75 | 1 |
| 9 | 8 | 3.75 | 5.17 | 3.75 | 1 |
| 10 | 0 | 4.25 | 2.83 | 2.83 | 2 |

Ahora se calcula la distancia de cada elemento a los nuevos centros, como se puede ver en la tabla 4. Este proceso se repite iterativamente hasta un número de veces propuesto por el usuario o hasta que no varié la configuración dentro de los conglomerados.

### 5.2.4.1.2 PAM (Partitioning around medoids)

Este método es parecido en cierto sentido a k-means. También trata de minimizar una función objetivo: minimizar las sumas de las distancias, pero es mucho más costoso, pues pasa por cada posible n-tupla de medias, compara sus distancias con todas las demás sumas de distancias, y escoge la menor; luego asigna elementos a este conglomerado. Como se aprecia, el concepto de ***centroide*** usado en k-means se convierte aquí en ***medoides***, no en el sentido de la mediana de los datos, sino en el sentido de que los centros forman parte de los datos. Entonces para un pre-especificado número de clusters K, el procedimiento busca los K medoides, M = ($m_1$,. . ., $m_K$) de todas las observaciones

a clasificar. Para encontrar M hay que minimizar la suma de las distancias de las observaciones a su más cercano medoide [10].

$$M^* = \arg\min_M \sum_i \min_k d(x_i, m_k) \quad (5.7)$$

donde $d$ es una medida de disimilaridad, $x_i$ es un elemento y $m_k$ un posible medodie
La función objetivo consiste en minimizar la suma de distancias a los K medoides. Esto matemáticamente es equivalente a minimizar el promedio de las distancias.
El siguiente ejemplo ilustra el algoritmo utilizado para escoger los medoides y asignar elementos a los conglomerados.

*Ejemplo en una dimensión:* Dado un conjunto de datos {1, 2, 4, 5, 8,10, 15,25} se quieren obtener dos conglomerados.

a) Supóngase que se inicia con centros iniciales localizados en las posiciones 1, 5
b) Se suman las distancias de cada punto a éstos centros inicales.

**Tabla 6.** Datos para el ejemplo de PAM

| ***Número*** | ***objeto*** | ***dist 1*** | ***dist 2*** | ***mínima d.*** | ***cluster*** |
|---|---|---|---|---|---|
| ***1*** | ***4*** | 0 | 6 | 0 | 1 |
| 2 | 1 | 3 | 9 | 3 | 1 |
| 3 | 2 | 2 | 8 | 2 | 1 |
| 4 | 5 | 1 | 5 | 1 | 1 |
| ***5*** | ***10*** | 6 | 0 | 0 | 2 |
| 6 | 15 | 11 | 5 | 5 | 2 |
| 7 | 8 | 4 | 2 | 2 | 2 |
| 8 | 25 | 21 | 15 | 15 | 2 |
| 9 | 8 | 4 | 2 | 2 | 2 |
| | | Average | | **3.333333** | |

Ahora supóngase que se toma como centros iniciales los elementos de las posiciones 2 y 8

**Tabla 7.** Nueva iteración de PAM

| ***Número*** | ***objeto*** | ***dist 1*** | ***dist 2*** | ***mínima d.*** | ***cluster*** |
|---|---|---|---|---|---|
| 1 | 4 | 3 | 21 | 3 | 1 |
| ***2*** | ***1*** | 0 | 24 | 0 | 1 |
| 3 | 2 | 1 | 23 | 1 | 1 |
| 4 | 5 | 4 | 20 | 4 | 1 |
| 5 | 10 | 9 | 15 | 9 | 1 |
| 6 | 15 | 14 | 10 | 10 | 2 |
| 7 | 8 | 7 | 17 | 7 | 1 |
| ***8*** | ***25*** | 24 | 0 | 0 | 2 |
| 9 | 8 | 7 | 17 | 7 | 1 |
| | | Average | | **4.555556** | |

El proceso ilustrado en las tablas 5 y 6 se repite con cada par de centros iniciales y finalmente se escoge el que menor suma o promedio de distancia tenga. En este ejemplo con sólo dos conglomerados se escogería la primera partición, pues es la que tiene menor promedio de mínima distancia, pero en realidad hace falta pasar por todas las posibles combinaciones de parejas, para luego escoger la que tenga menor suma o promedio de distancias a la pareja escogida. Se le llama mediodes.

### 5.2.4.2 Métodos Jerárquicos

Dada una matriz de distancias o de similitudes se desea clasificar los elementos en una jerarquía. Los algoritmos existentes funcionan de manera que los elementos existentes son asignados sucesivamente a grupos, pero la asignación es irrevocable, es decir, una vez hecha no se cuestiona nunca más.

Estos algoritmos son de dos tipos: de aglomeración o de división.

#### 5.2.4.2.1 Algoritmos jerárquicos aglomerativos

Estos algoritmos producen una sucesión de conglomerados de tal manera que en cada paso el número de conglomerados va disminuyendo. Son algoritmos del tipo “botton up”. Inicialmente se empieza con conglomerados que consisten de un sólo elemento. Los conglomerados de un paso dado son obtenidos al combinar dos conglomerados del paso anterior. Se uso el algoritmo AGNES que está dentro de está categoría porque esta muy bien documentado y disponible en la librería cluster del paquete estadístico R.

##### 5.2.4.2.1.1 AGNES (Agglomerative Nesting)

Este algoritmo construye una jerarquía en forma de árbol que contiene implícitamente todos los valores de *k*, comenzando con *N* conglomerados y siguiendo con fusiones sucesivas hasta obtener un sólo conglomerado con todos los objetos [10]. Podríamos esquematizar el funcionamiento del algoritmo como sigue:

(i) Si *n* es el número de elementos se comienza con tantas clases como elementos. Las distancias entre clases es las distancias entre los elementos originales.

(ii) Seleccionar los dos elementos más próximos en la matriz de distancias y formar con ellos una clase.

(iii) Sustituir los dos elementos utilizados en (ii) para definir la clase por un nuevo elemento que represente la clase construida. Las distancias entre este nuevo

elemento y los anteriores se calculan con cualquiera de los criterios que comentamos a continuación.

(iv) Volver a (2) y (3) hasta que tengamos todos los elementos agrupados en un una sola clase.

(v) Cortar el árbol donde se considere conveniente.

Para el enfoque aglomerativo, hay diferentes medidas de proximidad entre conglomerados estás se derivan de varias estrategias de fusión. Estas son conocidas como: encadenamiento simple, encadenamiento completo, encadenamiento de media de grupos, encadenamiento de "ward", y encadenamiento del centroide [11]. A continuación se describen:

***Encadenamiento simple o vecino más próximo:*** la distancia entre dos conglomerados es la distancia entre sus dos puntos más próximos. Es decir:

$$D(A,B)=min\ d(i,j) \quad (5.8)$$

Para cualquier elemento $i \in A$ y $j \in B$

***Encadenamiento completo o vecino más lejano:*** la distancia entre dos conglomerados A y B se calcula como la distancia entre sus puntos más lejanos *i, j.*

$$D(A,B)=máx\ d(i,j) \quad (5.9)$$

Para cualquier elemento $i \in A$ y $j \in B$.

Sea C el resultado de fusionar los conglomerados A y B. Se define $n_a$ y $n_b$ como el número de elementos de A y B respectivamente; $d_{CA}$ y $d_{CB}$ como las distancias del conglomerado C a los conglomerados A y B respectivamente.

***Encadenamiento de media de grupos*:** La distancia entre dos grupos es la media ponderada de las distancias entre grupos antes de la fusión, es decir, es el promedio de todas las distancias entre los objetos del conglomerado A y B.

$$d(C;AB) = \frac{n_a}{n_a + n_b} d_{CA} + \frac{n_b}{n_a + n_b} d_{CB} \quad (5.10)$$

***Encadenamiento del centroide:*** La distancia entre dos conglomerados es la distancia entre los centroides de los dos conglomerados:

$$d(C;AB) = \frac{n_a}{n_a + n_b} d^2{}_{CA} + \frac{n_b}{n_a + n_b} d_{CB} + \frac{n_a n_b}{(n_a + n_b)^2} d^2{}_{AB} \quad (5.11)$$

***Encadenamiento de ward:*** Para cada conglomerado, se calculan las medias para todas las variables. Después, para cada objeto, se calcula la distancia euclidiana cuadrada a las medias de los grupos; estás distancias se suman a todos los objetos es decir sus desviaciones cuadráticas. En cada etapa, se combinan los dos conglomerados con el menor incremento en la suma total de los cuadrados sus distancias.

$$W = \sum_{g} \sum_{i \in g} (X_{ig} - \overline{X_g})'(X_{ig} - \overline{X_g}) \quad (5.12)$$

Donde $g$ es un vector de conglomerados, $\overline{X_g}$ es el vector que contiene la media de cada conglomerado y $X_{ig}$, $i \in g$, es un vector que contiene los elementos del conglomerado $i$.

*Ejemplo del funcionamiento del AGNES:* Supóngase que se tiene la siguiente matriz de distancias, de un conjunto de cuatro elementos con dos variables, y utilizando la métrica euclidiana.

| | *a* | *b* | *c* | *d* | *e* |
|---|---|---|---|---|---|
| *a* | 0 | 2 | 6 | 10 | 9 |
| *b* | 6 | 0 | 5 | 9 | 8 |
| *c* | 6 | 5 | 0 | 4 | 5 |
| *d* | 10 | 9 | 4 | 0 | 3 |
| *e* | 9 | 8 | 5 | 3 | 0 |

El primer paso consiste en escoger los dos objetos más cercanos o más similares, es decir, donde la distancia sea más pequeña y fusionarlos, excepto los de la diagonal. Cómo se puede ver, el más pequeño es dos, así que unimos a y b para formar el conglomerado {a, b}. En este primer paso se formaron los conglomerados {a, b}, {c}, {d}, {e}. En el siguiente paso se unen los dos más cercanos, pero ahora para medir las distancias no lo tenemos que hacer de objeto a objeto sino entre conglomerados. Se usará entonces el *Encadenamiento de media de grupos*, discutido anteriormente. La bibliografía consultada sugiere mejores resultados que los demás [11]. Las distancias de los nuevos conglomerados será:

$$d(\{a, b\},\{c\}) = \frac{1}{2}\left[d(a,c) + d(b,c)\right] = 5.5$$

$$d(\{a,b\},\{d\}) = \frac{1}{2}\left[d(a,d) + d(b,d)\right] = 9.5$$

$$d(\{a,b\},\{e\}) = \frac{1}{2}[d(a,e) + d(b,e)] = 8.5$$

Se puede construir una nueva matriz de distancias (disimiláridades) entre los cuatro conglomerados {a, b}, {c}, {d}, {e}. La matriz de distancia es:

| | {a,b} | {c} | {d} | {e} |
|---|---|---|---|---|
| {a, b} | 0 | 5.5 | 9.5 | 8.5 |
| {c} | 5.5 | 0 | 4 | 5 |
| {d} | 9.5 | 4 | 0 | 3 |
| {e} | 8.5 | 5 | 3 | 0 |

Continuando con el procedimiento se buscan los mas similares y puede verse que la distancia entre {d} y {e} es la entrada más pequeña de la matriz. Los cálculos de distancia para los nuevos conglomerados son:

$$d(\{d,e\},\{c\}) = \frac{1}{2}[d(d,c) + d(e,c)] = 4.5$$

$$d(\{d,e\},\{a,b\}) = \frac{1}{4}[d(d,a) + d(d,b) + d(e,a) + d(e,b)] = 9.0$$

De nuevo se tiene otra matriz de distancias ahora con {a,b},{c}, {d,e}, como se muestra a continuación.

| | {a,b} | {c} | {d,e} |
|---|---|---|---|
| {a,b} | 0 | 5.5 | 9.0 |
| {c} | 5.5 | 0 | 4.5 |
| {d,e} | 9.0 | 4.5 | 0 |

El termino más pequeño ahora es 4.5, luego podemos fusionar {d, e} y {c}, el calculo de distancias entre los conglomerados resultantes es:

$$d(\{c,d,e\},\{a,b\}) = \frac{1}{6}[d(c,a)+d(c,b)+d(d,a)+d(d,b)+d(e,a)+d(e,b)] = 7.83$$

Lo que conduce a la siguiente matriz de distancias como se muestra a continuación:

| | {a,b} | {c,d,e} |
|---|---|---|
| {a,b} | 0 | 7.83 |
| {c,d,e} | 7.83 | 0 |

El paso final consiste en unir los dos últimos conglomerados en uno sólo. La siguiente Figura ilustra el proceso de fusión que sufrieron los elementos.

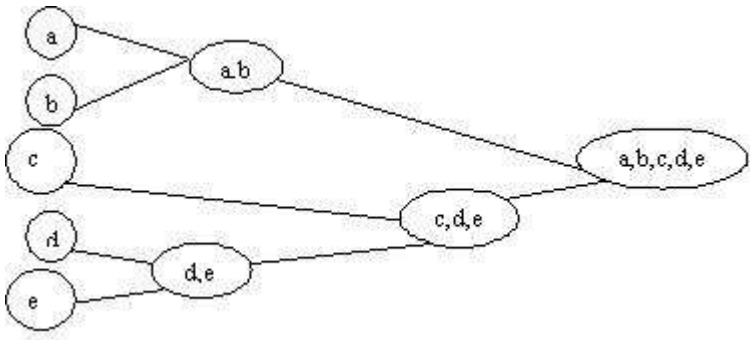


**Figura 6 Proceso de fusión con el algoritmo AGNES**

#### **5.2.4.2.2** Algoritmos jerárquicos divisivos

Al igual que los algoritmos aglomerativos, estos algoritmos, también producen una sucesión de conglomerados, pero en este caso el número de ellos crece en cada paso. Son algoritmos del tipo “top-down”. Inicialmente se empieza con un sólo conglomerado que contiene a todas las observaciones. Los conglomerados de un paso dado son obtenidos al dividir en dos un conglomerado del paso anterior. Los algoritmos jerárquicos divisivos demandan más esfuerzo computacional que los algoritmos aglomerativos. En está tesis se utilizo el algoritmo *DIANA* que pertenece a está categoría .

##### 5.2.4.2.2.1 DIANA (Divisive Analysis)

DIANA es un algoritmo jerárquico divisivo que comienza con un gran conglomerado que contiene a todos los *n* objetos. En cada etapa se selecciona el conglomerado con diámetro más grande. El diámetro de un conglomerado es la distancia mas grande entre cualquiera dos de sus objetos. Para dividir el conglomerado primero se busca su objeto más disimilar o distante, es decir, el que tiene la mayor distancia promedio a otros objetos del conglomerado. Con este objeto se inicia el particionado de grupos. En los pasos consecutivos, el algoritmo reasigna los objetos que son más cercanos a cada uno de los nuevos grupos. De está manera, elementos del grupo viejo pueden pasar al grupo nuevo, resultando en una división de él en dos nuevos.
Su algoritmo se describe a continuación:

*(i)* Calcule el *diámetro* de cada conglomerado Q

(ii) Seleccione el conglomerado con diámetro mas grande

i) Divídalo en A y B

(iii) Seleccione el objeto *i* del conglomerado A que maximice el promedio de las disimilaridades a todos los demás objetos de A.

(iv) Si la diferencia entre la distancia promedio de *i* a los demás elementos de A y la distancia promedio de *i* a los elementos de B es mayor que cero. Mueva i de A a B.

(v) Si no, vuelva a (i), hasta que hayan tantos conglomerados como elementos.

*Ejemplo numérico: Para* este ejemplo utilizaremos la misma matriz de disimiláridades usada en el ejemplo aglomerativo. Esta matriz se muestra a continuación:

| | *a* | *b* | *c* | *d* | *e* |
|---|---|---|---|---|---|
| *a* | 0 | 2 | 6 | 10 | 9 |
| *b* | 6 | 0 | 5 | 9 | 8 |
| *c* | 6 | 5 | 0 | 4 | 5 |
| *d* | 10 | 9 | 4 | 0 | 3 |
| *e* | 9 | 8 | 5 | 3 | 0 |

Supóngase que lo objetos *a, b, c, d* y e forman un sólo conglomerado. Esto no se hace considerando todas la posibles particiones, sino por una clase de procedimiento iterativo. El primer paso consiste en buscar el miembro más disímil o alejado a todos los demás, pero para hacer esto es necesario definir una medida de disimilitud entre un objeto y un grupo de objetos. El algoritmo usa el promedio de las disimilitudes (*average*

*dissimilarity*). Se busca el objeto para el cual el promedio de las distancias a todos los demás objetos sea más grande. Cuando hay dos de tales objetos escogemos uno al azar. En este ejemplo se obtiene la tabla siguiente:

| Objeto | Promedio de disimiláridades a otros objetos |
|---|---|
| *A* | (2+6+10+9)/4=6.75 |
| *B* | (2+5+9+8)/4=6 |
| *C* | (6+5+4+5)/4=6.5 |
| *D* | (10+9+4+3)/4=6.5 |
| *E* | (9+8+5+3)/6.25=6.25 |

De esta tabla se observa que el elemento con mayor disimiláridad es *a* y con el inicia el llamado "*splinter group"*. En está etapa se obtienen los conglomerados {*a*}, {*b, c, d, e*}. Luego, para cada objeto del grupo más grande se calcula la disimiláridad promedio a los objetos restántes y se compara su disimiláridad promedio con la del grupo particionador (en este caso *a*)

| Objetos | Promedio de disimiláridad de los objetos restantes | Promedio disimiláridad del grupo particionador | Diferencias |
|---|---|---|---|
| *b* | (5+9+8)/3 ≈ 7.33 | 2 | 5.33 |
| *c* | (5+4+3)/3 ≈ 4.67 | 6 | -1.33 |
| *d* | (9+4+3)/3 ≈ 5.33 | 10 | -4.67 |
| *e* | (8+5+3)/3 ≈ 5.33 | 9 | -3.67 |

La mayor diferencia corresponde a *b*. Así, este objeto cambia de conglomerado, y el nuevo grupo particionador es {*a, b*} y el grupo restánte {*c, d, e*}. Repitiendo los cálculos se obtiene la tabla siguiente:

| Objetos | Promedio de disimiláridad de los objetos restantes | Promedio disimiláridad del particionador | Diferencias |
|---|---|---|---|
| *c* | (4+5)/2=4.5 | (6+5)/2=5.5 | -1 |
| *d* | (4+3)/2=3.5 | (10+9)/2=9.5 | -6 |
| *e* | (5+3)/2=4 | (9+8)/2=8.5 | -4.5 |

Como no hubo ninguna diferencia positiva se detiene el proceso y concluye el primer paso divisivo, el cual partió el conglomerado inicial en dos conglomerados {*a, b*}, {*c, d, e*}. En el siguiente paso se divide el conglomerado más grande, es decir el que tenga el mayor diámetro. El diámetro de {a, b} es 2 y el de {c, d, e} es 5 como se ve en la siguiente matriz

| | *c* | *d* | *e* |
|---|---|---|---|
| *c* | 0 | 4 | 5 |
| *d* | 4 | 0 | 3 |
| *e* | 5 | 3 | 0 |

Para ver cuál será el próximo elemento que abandone el conglomerado, realizan los mismos cálculos anteriores en el grupo con mayor diámetro y se encuentra que *c* tiene el mayor promedio de disimilaridad.

| Objetos | Promedio de disimiláridad los demás objetos |
|---|---|
| *c* | (4+5)/2=4.5 |
| *d* | (4+3)/2=3.5 |
| *e* | (5+3)/2=4 |

Comparando con el grupo particionador se observa en la tabla siguiente que ninguna diferencia resultó positiva.

| Objetos | Promedio de disimiláridad de los objetos restántes | Promedio disimiláridad del grupo particionador | Diferencias |
|---|---|---|---|
| *d* | 3 | 4 | -1 |
| *e* | 3 | 5 | -2 |

De nuevo se suspende el proceso. El conglomerado {c, d, e} quedó dividido en {c}, {d, e}. Se tiene entonces los conglomerados {a, b}, {c} y {d, e}. Ahora se debe decidir cuál

de estos partir. El conglomerado {a, b} tiene diámetro 2 y {d, e} tiene diámetro 3, así que siguiendo el algoritmo, se divide el conglomerado de mayor diametro {d, e}

| | *d* | *e* |
|---|---|---|
| *d* | 0 | 3 |
| *e* | 3 | 0 |

Para comenzar la partición, calculamos la disimilaridad de cada objeto de {d, e} dentro del mismo conglomerado.

| Objetos | Promedio de disimiláridad los demás objetos |
|---|---|
| *d* | 3 |
| *e* | 3 |

Como el promedio de disimiláridad es igual, se puede escoger cualquiera de los dos para comenzar el grupo particionador. Escoja por ejemplo el objeto *d*. Así se obtiene {d}, {e}. Como el objeto *e* es el último que queda no puede unirse al grupo particionador. Después del tercer paso quedan los conglomerados: {a, b}, {c},{d}, {e}. Cómo en el tercer paso, se divide {a, b} en {a},{b}. Después del cuarto paso se obtienen los conglomerados {a},{b},{c},{d},{e} y el algoritmo para. La figura 6 ilustra el proceso.

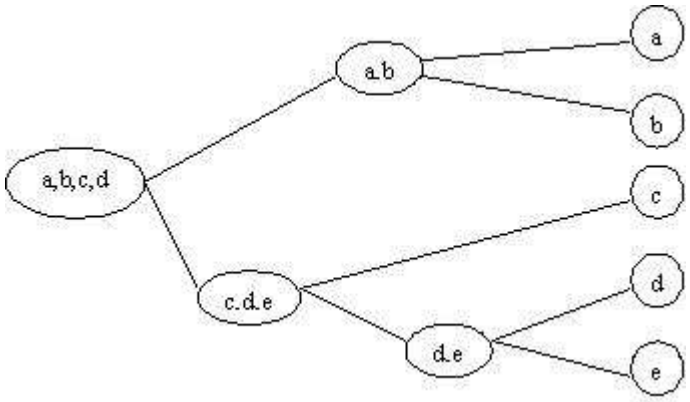


**Figura 6  Proceso de partición usando DIANA**

# CAPÍTULO 6: PANORAMA GENERAL DE LAS MEDIDAS PARA EVALUAR LA CALIDAD DE LOS CONGLOMERADOS

Muchos algoritmos de agrupamiento generan dos o más particiones de los datos. Para validar éstas debe darse respuestá a las preguntas: ¿Cuántos grupos hay en los datos? o ¿Dónde debe cortarse un dendrograma?. En dicha validación son útiles los criterios para interrumpir la formación de los grupos o reglas de interrupción (*stopping rules*), en las que se produce una secuencia de cortes del dendrograma en niveles sucesivos; con las particiones así obtenidas se aplica una regla de decisión para determinar el mejor nivel en el que debe cortarse la formación de grupos.

Milligan y Cooper (1985) efectuaron un estudio comparativo mediante simulación Monte Carlo que incluyó 30 de esas reglas de interrupción. Concluyeron que entre las mejores están los índices CA(K) de Calinski-Harabasz, y DB(K) de Davies-Bouldin. Estás reglas son utilizadas en este proyecto, junto con enfoques más recientes como el Average *Silhouette* propuesto por Rousseeuw en 1987 y el *Mean Split Silhouette* propuestos por *Pollard & van der Laan en 2002.* A continuación se describe brevemente cada una de estás medidas de validación.

## 6.1 Average Silhouette

Está medida fue introducida por Rousseeuw (1987) y puede ser usada para:

(i) Seleccionar el número de conglomerados.

(ii) Evaluar cuan bien han sido asignadas las observaciones a los conglomerados.

El ancho de la silueta está definido por:

$$sil_i = (b_i - a_i)/ \text{máx.}(a_i, b_i) \quad (6.1)$$

dónde $a_i$ denota la distancia promedio entre la observación *i* y todas las otras que están en el mismo conglomerado de *i*, y $b_i$ denota la distancia promedio mínima de *i* a las observaciones que están en otros conglomerados. Intuitivamente, observaciones con ancho de silueta grande están bien agrupadas mientras aquellas con ancho de silueta baja tienden a estár ubicadas en el medio de dos conglomerados [1].

Para un número de conglomerados dado, *K*, el ancho de silueta promedio de la configuración de conglomerados será simplemente el promedio de $sil_i$ sobre todas las observaciones. Es decir

$$\bar{s} = \frac{\sum_i sil_i}{n} \quad (6.2)$$

Kaufman y Rousseeuw (1990) sugirieron estimar el número óptimo de conglomerados K para el cual el promedio del ancho de silueta $\bar{s}$ es el mayor posible.

*Ejemplo en una dimensión:* Dado el conjunto A = {1,2,3,7,8,9}, con dos conglomerados: {1, 2, 3} y {7, 8, 9}, las siluetas de cada observación son las siguientes:

$$a(1) = \frac{1+2}{2} = 1.5 \qquad b(1) = \frac{6+7+9}{3} = 7 \quad s(1) = \frac{7-1.5}{7} = 0.7857143$$

$$a(2) = \frac{1+1}{2} = 1 \qquad b(2) = \frac{5+6+7}{3} = 6 \quad s(2) = \frac{6-1}{6} = 0.8333333$$

$$a(3) = \frac{1+2}{2} = 1.5 \qquad b(3) = \frac{4+5+6}{3} = 5 \quad s(3) = \frac{5-1.5}{5} = 0.7$$

Así, el promedio de las siluetas del primer conglomerado es: 0.7730159. De igual forma lo hacemos con cada uno de los elementos de del segundo conglomerado:

$$a(1) = \frac{1+2}{2} = 1.5 \qquad b(1) = \frac{6+7+9}{3} = 7 \quad s(1) = \frac{7-1.5}{7} = 0.7857143$$

$$a(2) = \frac{1+1}{2} = 1 \qquad b(2) = \frac{5+6+7}{3} = 6 \quad s(2) = \frac{6-1}{6} = 0.8333333$$

$$a(3) = \frac{1+2}{2} = 1.5 \qquad b(3) = \frac{4+5+6}{3} = 5 \quad s(3) = \frac{5-1.5}{5} = 0.7$$

En este caso los datos son similares a los de primer conglomerado y su promedio es: 0.7730159. El promedio del ancho de la *Silhouette* de la partición *k=2* es: $\frac{0.7730159 + 0.7730159}{2} = 0.7730159$

## 6.2 Mean Split Silhouette

Dados K grupos de conglomerados, considérese cada grupo k = 1, . . . , K separadamente [12].

• Escoja el número de hijos que va a tener cada conglomerado del nivel K.

• Aplique el algoritmo de agrupación a los elementos de los conglomerados del nivel K independientemente de los elementos en otros niveles

•Evalue la función objetivo sobre cada uno de los de los conglomerados en cada nivel K. Para esto se escoge un número de hijos en cada nivel de conglomerados que maximiza el promedio del ancho de la silueta (*average silhouette width*). A este promedio se lo conoce como el *mean split silhouette*. El mínimo valor de la silueta entre todos los niveles de *k* indica la partición de conglomerados, donde la mayoría de ellos son más homogéneos. La idea detrás de éste método es evaluar qué tan bien los elementos en un

conglomerado pertenecen a este. Esto se hace mediante la aplicación del algoritmo y la función objetivo sólo a un nivel K, ignorado los demás niveles.

## **6.3** Coeficiente de Calinski

Este coeficiente se basa en la suma de cuadrados dentro (*SSW*) y entre (*SSB*) los conglomerados. Está medida de dispersión dentro y entre se define de la siguiente manera [13]:

$$Calinski = (SSB/(k-1))/(SSW/(n-k)) \quad (6.3)$$

donde *n* es el número de datos y *k* el número de conglomerados

$$SSW = \left\{ \frac{1}{N_i} \sum_{j=1}^{N_i} \left| x_j - c_i \right|^2 \right\}^{1/2} \quad (6.4)$$

es la suma de cuadrados de las distancias dentro de los conglomerados

$$SSB = \left\{ \sum_{i=1}^{d} \left| c_{ki} - c_{kj} \right|^2 \right\}^{1/2} \quad (6.5)$$

es la distancia cuadrada entre los conglomerados, medida entre sus respectivos centroides. Este índice mide la dispersión de los datos dentro y entre los cluster. El siguiente ejemplo ilustra cómo trabaja este índice:

Sea A = {(1, 1,), (3, 1), (6, 1), (8, 1)} un conjunto de puntos en dos dimensiones. Si se agrupan en dos conglomeraos, se tienen los conglomerados $k_1$={(1, 1), (3, 1)} y $k_2$={(6, 1), (8, 1)} y sus respectivos centroides {(2, 1), (7, 1)}. Calculando *SSB* Y *SSW* obtenemos

$$SSB=\left(\sqrt{(7-2)^2+(1-1)^2}\right)^2=25$$

$$SSW=\frac{\left(\sqrt{(3-1)^2+(1-1)^2}\right)^2+\left(\sqrt{(8-6)^2+(1-1)^2}\right)^2}{2}=4$$

donde *k* es el número de conglomerados y *n* el número de elementos. Entonces tenemos que para *k=2 y n=4.*

$$Coeficiente\ de\ Calinski=\frac{\frac{25}{2-1}}{\frac{4}{4-2}}=12.5$$

## 6.4 Coeficiente de DB

El índice Davies-Bouldin [13] es una medida que indica lo similares pueden ser dos agrupaciones. Está medida puede ser usada para validar la partición, es decir, para comparar las diferentes particiones del conjunto de datos. Este índice es independiente del número de conglomerados o del algoritmo usado para hacer la partición. Formalmente podemos decir que es una función del cociente de la suma de la dispersión de dentro del conglomerado sobre la dispersión entre-conglomerados. Este coeficiente está dado por:

$$DB\text{-}Index=(1/n)\sum_{i=1}^{n}R_{ij} \quad (6.6)$$

Donde $R_{ij}= (SSW_i + SSW_j)/DC_{ij,}$ *y* $DC_{ij}$ es la distancia entre los centros de los conglomerados i, j, para $i \neq j$.

$SSW_i$ representa la distancia intra-conglomerado, es decir el diámetro del conglomerado y $c_i$ representa el centroide del conglomerado. Así $SSW_i$ está dado por la formula (6.4), donde $x_j \in W_k$, $N_k$ es el número de muestras en el conglomerado $W_k$, $c_i = 1/N_i \sum_{x_{t \in W_i}} x_t$ y $DC_{kl}$ *o* $dc(W_i, W_j)$, es la distancia euclidiana entre los centroides y se define como:

$$DC_{kl=} \left\{ \sum_{i=1}^{d} \left| c_{ki} - c_{kj} \right|^2 \right\}^{1/2} \quad (6.7)$$

Donde *d* corresponde a la dimensión del vector $x_k$, El mínimo valor es tomado como el número adecuado de conglomerados. El siguiente ejemplo ilustra el funcionamiento del algoritmo.

Sea A = {(1, 1,), (3, 1), (6, 1), (8, 1)} un conjunto de puntos en dos dimensiones. Si se agrupa en dos conglomerados obtenemos $k_1$={(1, 1), (3, 1)} y $k_2$={(6, 1), (8, 1)} y sus respectivos centroides son: {(2, 1), (7, 1)}. Calculando el índice de *db* para está agrupación obtenemos:

$$\text{DB-Index} = \frac{1}{2} * \frac{(2+2)}{5} = 0.4$$

donde 2 es el número de agrupaciones, 2 es el diámetro del primer conglomerado, 2 es el diámetro del segundo y 5 es la distancia entre los conglomerados midiéndola entre sus centros, es decir, dist{(2, 1), (7, 1)}=5.

# CAPÍTULO 7: SIMULACIONES

Para desarrollar la metodología, se compararon los cuatro métodos de agrupación mencionados en el Capítulo 5 y las cuatro medidas validación para determinar la calidad de los conglomerados mencionadas en el Capítulo 6. El análisis se realizó con los siguientes datos simulados.

## 7.1 Datos de las simulaciones

Se diseñó una estructura de datos que es una idealización de la estructura que se espera encontrar en las observaciones con datos educativos. Estos datos se generaron con una distribución normal multivariable con cuatro, seis, ocho, diez, quince y veinte variables. En cada conjunto se generaron los datos para que hubiera, tres, cinco, siete y nueve conglomerados. Los conglomerados difieren en sus medias y estás medias se han espaciado uniformemente. Los elementos no están correlacionados y comparten el valor de la desviación estándar. Gradualmente se aumenta el valor de la desviación, se hace menos clara la distinción entre conglomerados, lo que se considera normal pues cuando el valor de la desviación estándar aumenta comienza haber solapamiento con los intervalos de confianza de medias cercanas.

Los conglomerados son de igual tamaño y cada uno de ellos contiene 25 datos normalmente distribuidos alrededor de cada una de las medias. Además se consideró el

siguiente rango de valores para la desviación estándar *sd* = {0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.2} y se varió la dimensión de las observaciones, con los siguientes valores de *p*={4, 6, 8,10, 15} y {2,4,6},{2,4,6,8,10},{2,4,6,8,10,12,14}, {2,4,6,8,10,12,14,16,18} las medias de los conglomerados

## 7.2 Metodología usada en las simulaciones

Para obtener una muestra del comportamiento de los algoritmos con está estructura de datos, se hicieron veinticinco simulaciones por tripla (*desviación estándar, dimensión de los datos, número de conglomerados*). Por ejemplo para datos con *p* =4, se mantuvo fijo el número de conglomerados, *nc* (número de conglomerados)=3 y se varió *sd*. Entonces se hicieron 25 simulaciones con *p* = 4, *nc* = 3, *sd* = 0.5, y otras 25 con *p* = 4, *nc* = 3 y *sd* = 0.6 y así sucesivamente, para *p* = {4, 6, 8,10, 15}, *nc* = {3, 5, 7, 9} y *sd* = {0.5, 0.6, 0.7, 0.8, 0.9, 1, 1.2}. En cada simulación se permutaron los datos antes de aplicar k-means y a los conglomerados resultantes de aplicar todos algoritmos de agrupación, se los midió con las medidas de validación. Este proceso dio como resultado una serie de gráficas en la cuales el eje *x* corresponde a la desviación estándar y el eje *y* al número de aciertos que en 25 oportunidades tuvo el método-medida, en términos de porcentaje. Los significados de las siglas que se usaron en las gráficas pueden verse en la Tabla 2.

Las combinaciones restántes no fueron incluidas pues su porcentaje de acierto fue muy bajo. Además se desarrollo el código para que el coeficiente de calinski pudiera medir los conglomerados agrupados por PAM pues sólo estába desarrollado para medir los conglomerados agrupados por k-means. Igualmente se hizo con el coeficiente de DB. Por otra parte hay que aclarar que el algoritmo k-means se corrió dejando que los centroides se escogieran al azar, como está implementado por defecto en el paquete estadístico R. Es necesario actualizar algunas de las bibliotecas de este paquete estadístico para que algunas de las medidas de validación sean compatibles con todos los algoritmos de agrupación.

Las Figuras 8 a la 30 son el resultado de aplicar la metodología anteriormente descrita a los conjuntos de datos mencionados anteriormente.

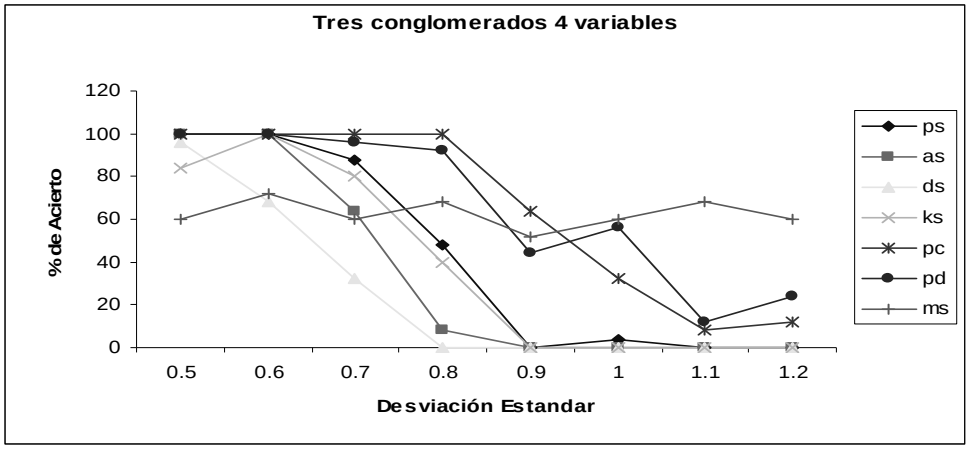


**Figura 8.** Porcentaje de acierto. Tres conglomerados y datos con cuatro variables

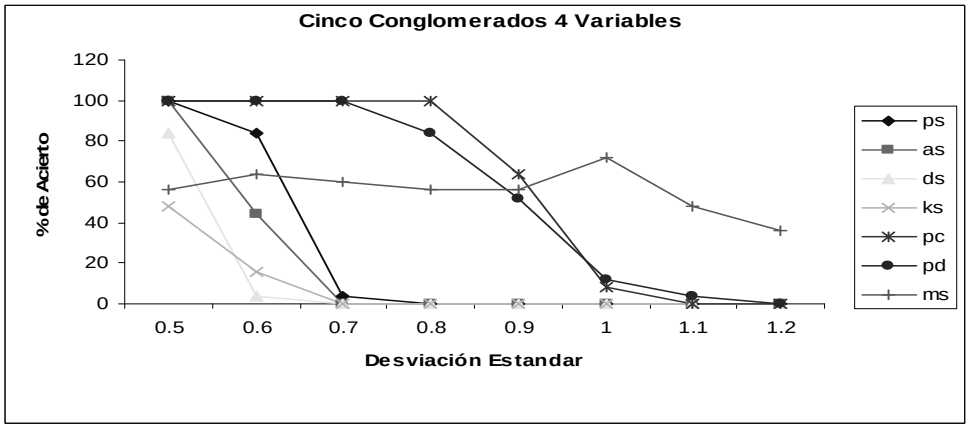


**Figura 9.** Porcentaje de acierto. Cinco conglomerados y datos con cuatro variables

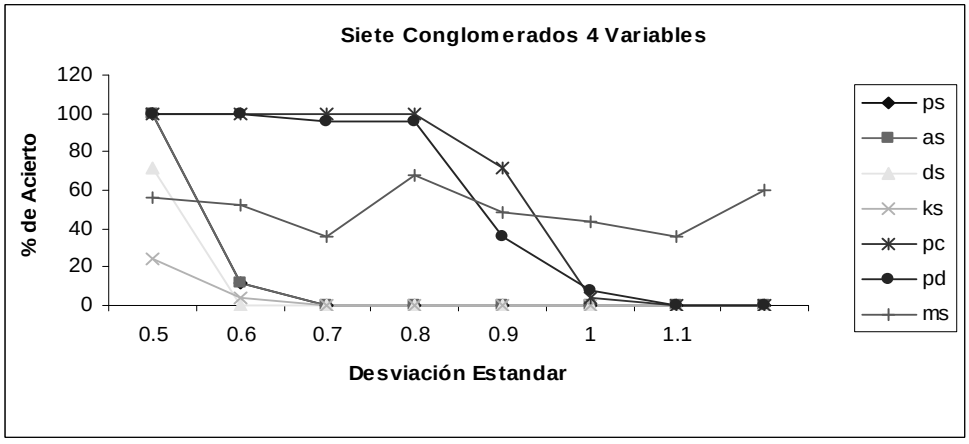


**Figura 10.** Porcentaje de acierto. Siete conglomerados y datos con cuatro variables

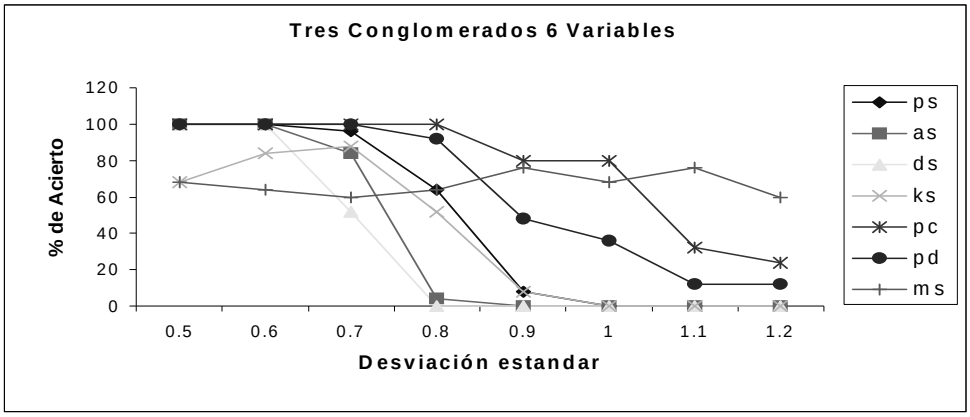


**Figura 11.** Porcentaje de acierto. Tres conglomerados datos con seis variables

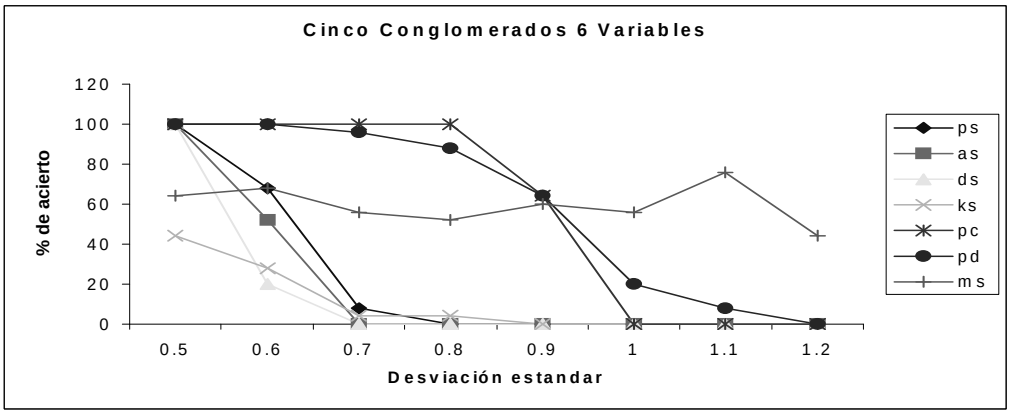


**Figura 12.** Porcentaje de acierto. Cinco conglomerados y datos con seis variables

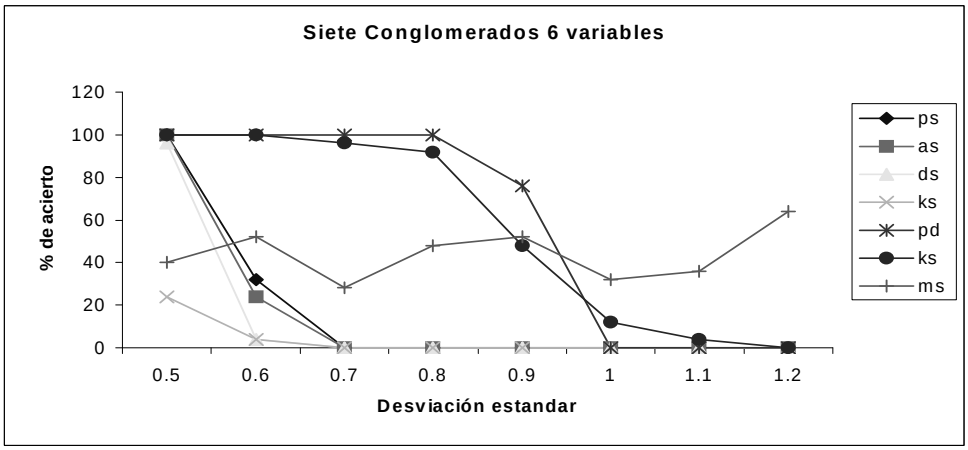


**Figura 13.** Porcentaje de acierto. Siete conglomerados y datos con seis variables

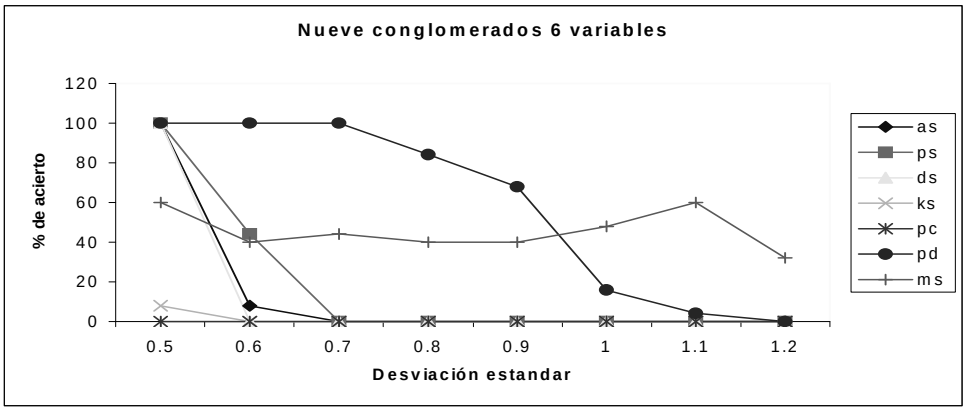


**Figura 14.** Porcentaje de acierto. Nueve conglomerados y datos con seis variables

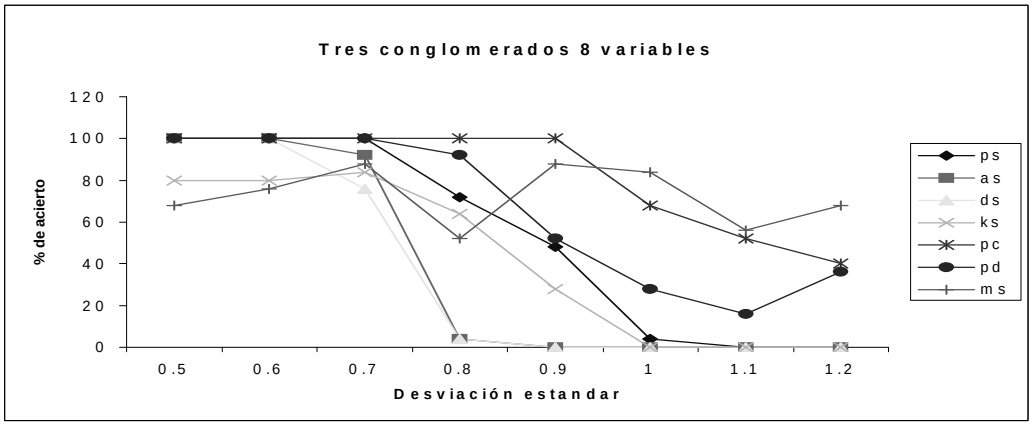


**Figura 15.** Porcentaje de acierto. Tres conglomerados y datos con ocho variables

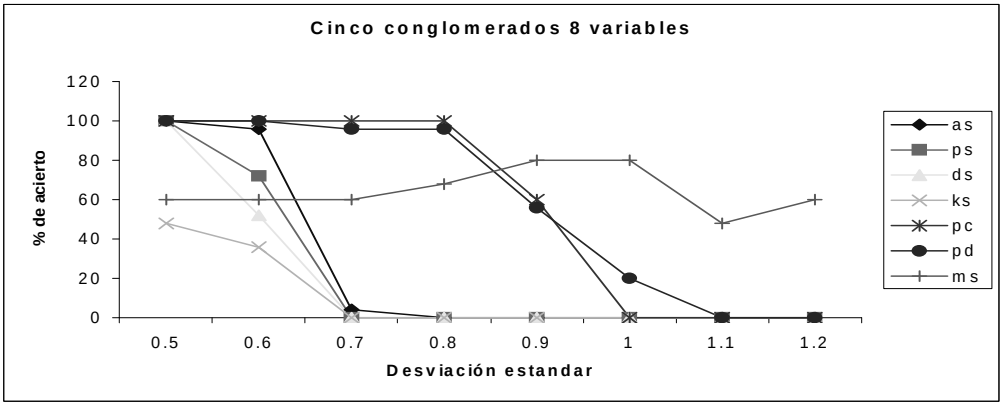


**Figura 16.** Porcentaje de acierto. Cinco conglomerados y datos con ocho variables

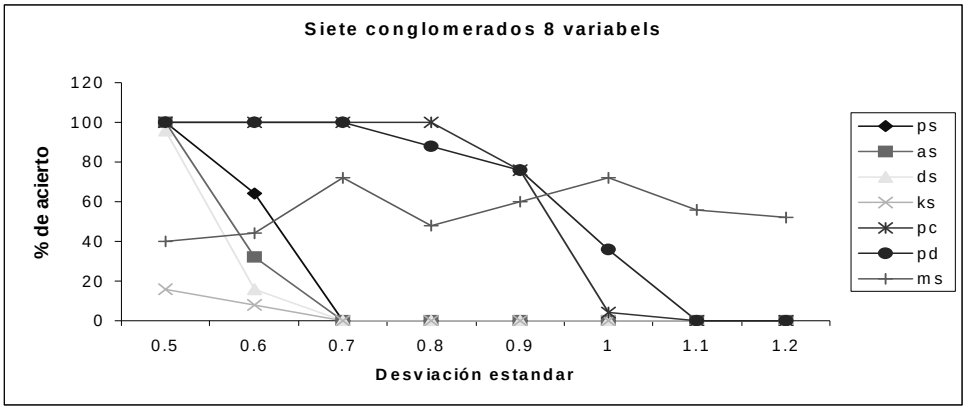


**Figura 17.** Porcentaje de acierto. Siete conglomerados y datos con ocho variables

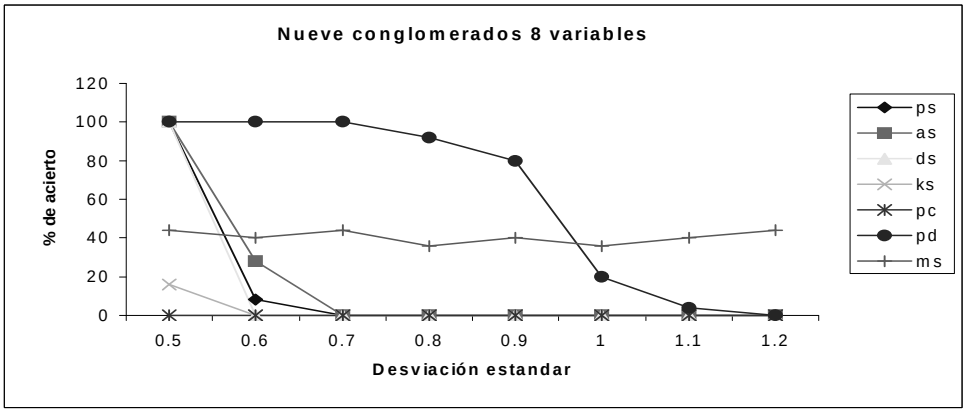


**Figura 18.** Porcentaje de acierto. Nueve conglomerados y datos con ocho variables

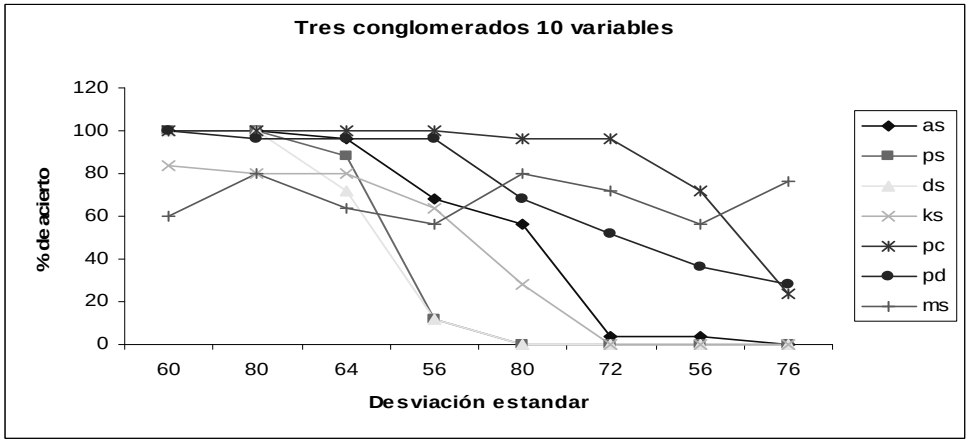


**Figura 19.** Porcentaje de acierto. Tres conglomerados y datos con diez variables

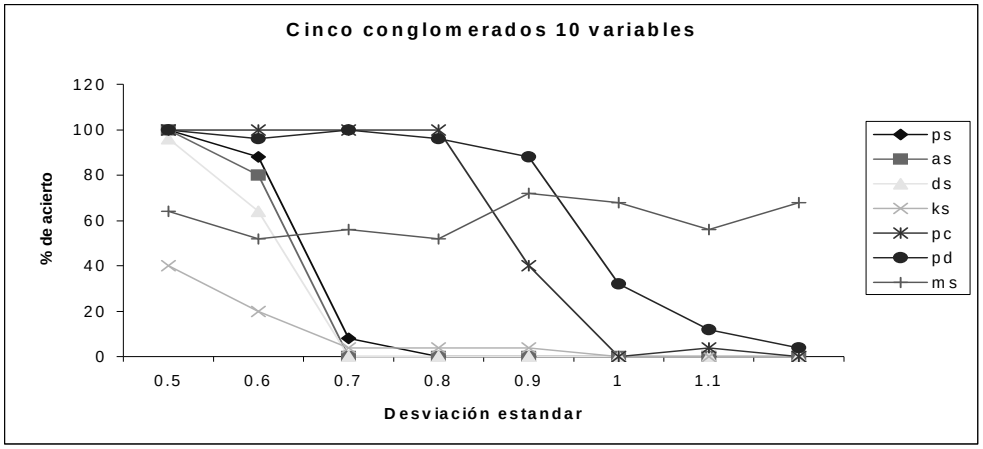


**Figura 20** Porcentaje de acierto. Cinco conglomerados y datos con diez variables

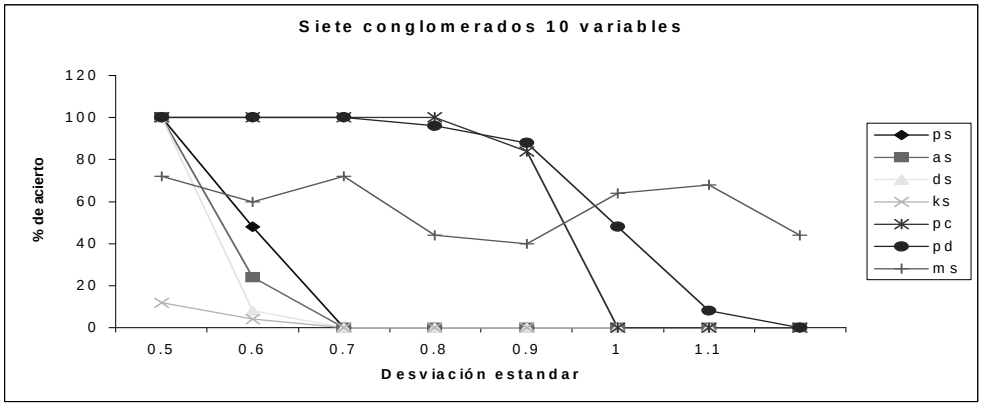


**Figura 21** Porcentaje de acierto. Cinco conglomerados y datos con diez variables

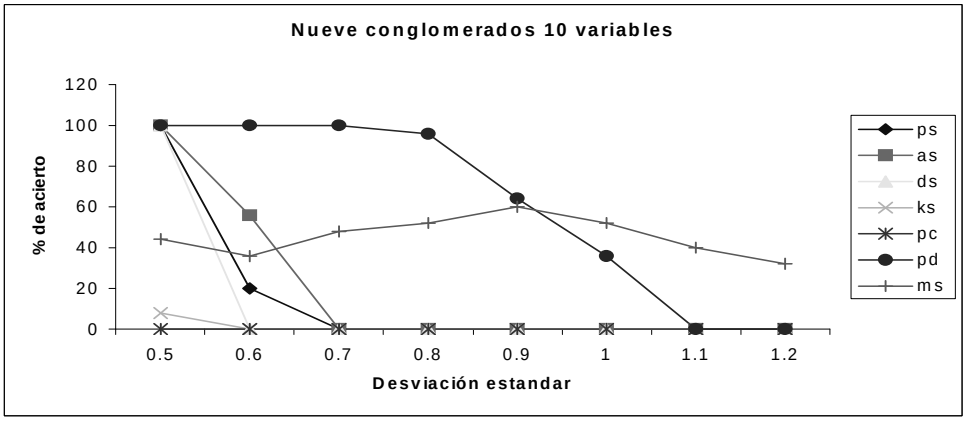


**Figura 22** Porcentaje de acierto. Nueve conglomerados y datos con diez variables

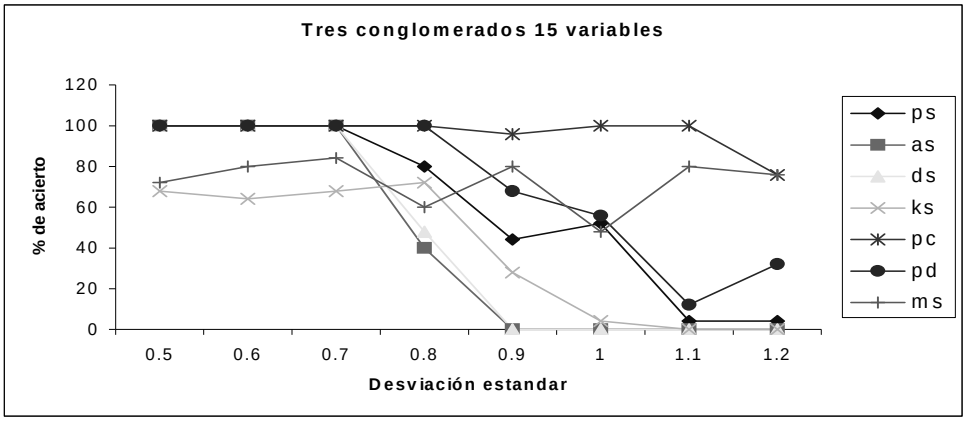


**Figura 23** Porcentaje de acierto. Tres conglomerados y datos con quince variables

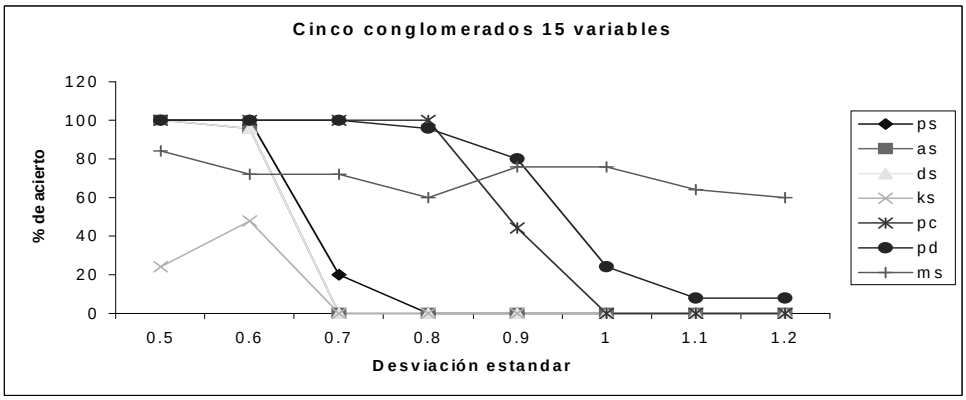


**Figura 24** Porcentaje de acierto. Cinco conglomerados y datos con quince variables

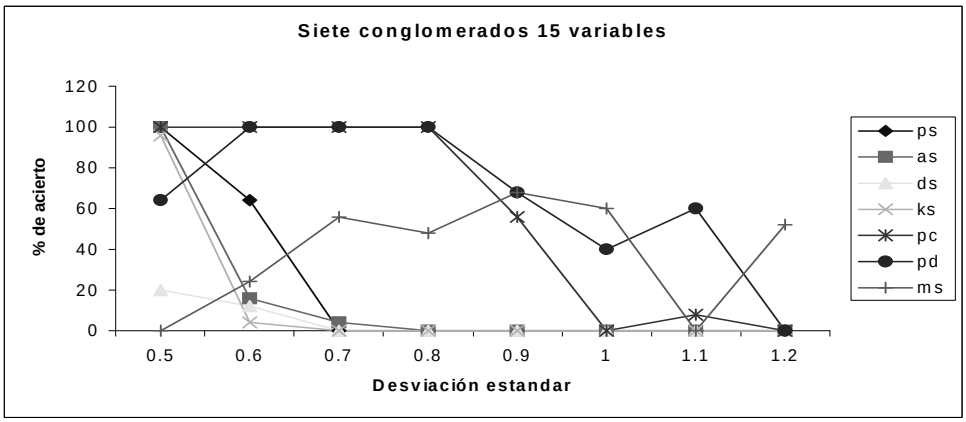


**Figura 25** Porcentaje de acierto. Siete conglomerados y datos con quince variables

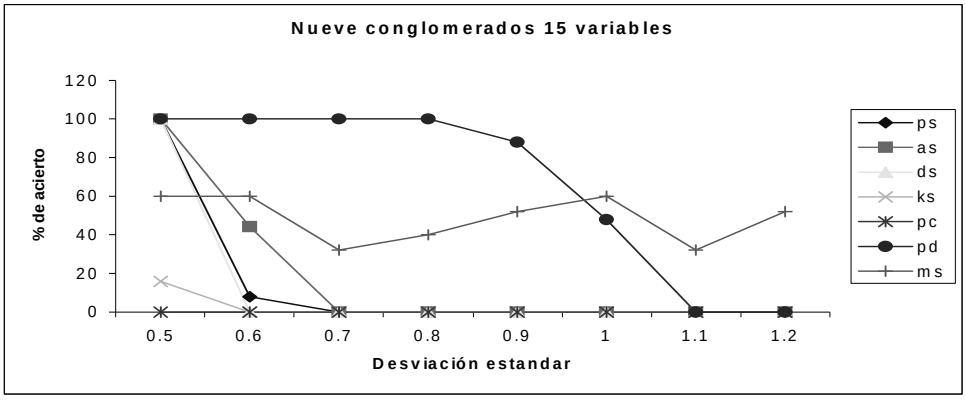


**Figura 26** Porcentaje de acierto. Nueve conglomerados y datos con quince variables

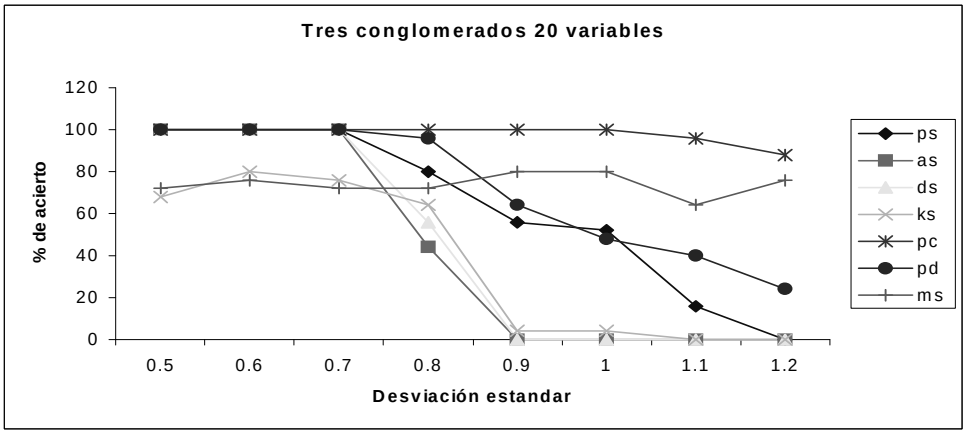


**Figura 27** Porcentaje de acierto. Tres conglomerados y datos con veinte variables

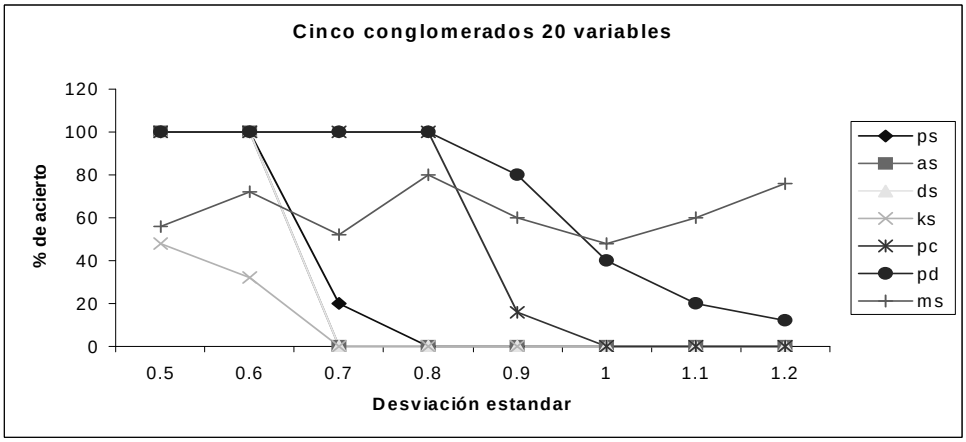


**Figura 28** Porcentaje de acierto. Cinco conglomerados y datos con veinte variables

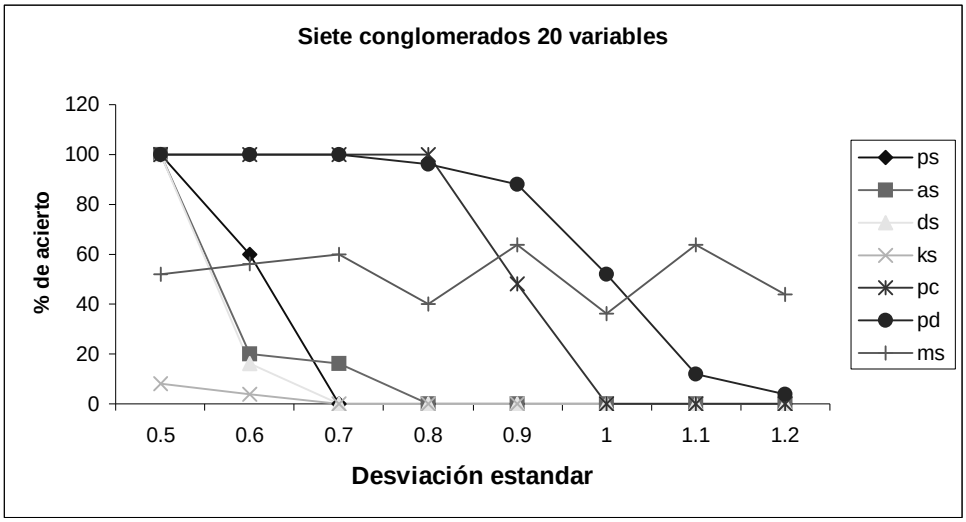


**Figura 29** Porcentaje de acierto. Siete conglomerados y datos con veinte variables

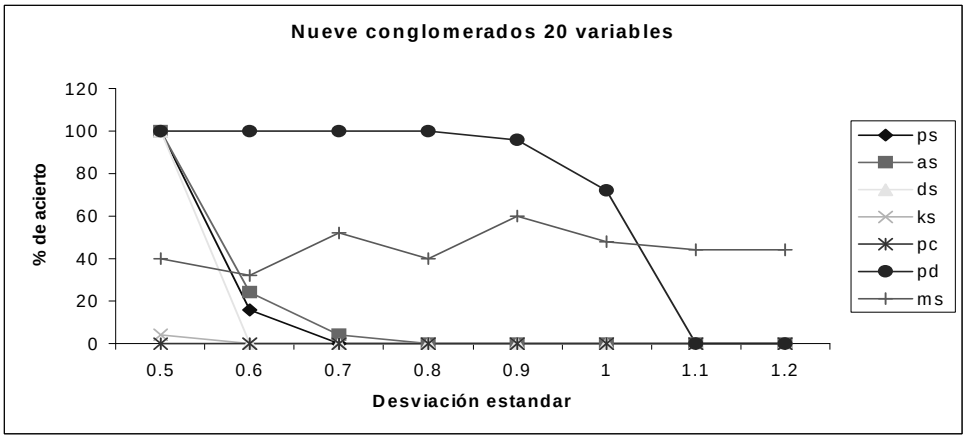


**Figura 30** Porcentaje de acierto. Nueve conglomerados y datos con veinte variables

## 7.4 Resultados de las simulaciones

Se puede ver que con centros separados a dos unidades de distancia, y con una desviación estándar de 0.5, la mayoría de las medidas de validación identificaron los

conglomerados, excepto k-means/Silhouette, resultado que se considera razonable pues de dejo que k-means tomara los datos iniciales aleatoriamente.

Se encontró que con centros separados a dos unidades de distancia, se perdió la capacidad de identificar los conglomerados en las agrupaciones hechas por Pam/Mean Split Silhouette, a partir de una desviación estándar 1.2, aunque no perdió su capacidad de identificar conglomerados hasta ese valor de desviación estándar. Sin embargo no fue tan bueno con desviaciones menores, lo que indica que es un buen indicador para estructuras anidadas y tiende a buscar conglomerados muy pequeños. Sin embargo PAM/DB y PAM/Calinski con desviaciones estándar de hasta 0.85 y en ocasiones hasta 1 mostraron un 100 % de acierto en el número correcto de conglomerados, lo que los convierte en muy buenos indicadores.

Las simulaciones muestran que con centros separados a dos unidades de distancia, se perdió totalmente la capacidad de identificar los conglomerados en las agrupaciones hechas por k-means/ Silhouette, a partir de una desviación estándar de 0.74

Se observa a través de las simulaciones que con centros separados a dos unidades de distancia, se perdió totalmente la capacidad de identificar los conglomerados en las agrupaciones hechas por AGNES/ Silhouette, a partir de una desviación estándar 0.76

Se encontró que con centros separados a dos unidades de distancia, se perdió totalmente la capacidad de identificar los conglomerados en las agrupaciones hechas por DIANA/ Silhouette, a partir de una desviación estándar 0.74

En general, las simulaciones muestran que los centroides de los conglomerados deben estar separados uno del otro a una distancia de aproximadamente tres veces la desviación estándar de los datos, para que todos los métodos-medida los identifiquen correctamente.

# CAPÍTULO 8: RESULTADOS.

## 8.1 Resumen de los datos y los métodos usados en la tercera etapa

En está etapa se busca identificar los conglomerados que se manifiestan en todos los métodos y en diferentes agrupaciones simultáneamente, utilizando los datos reales. Por ejemplo, si agrupamos en cinco conglomerados, utilizando un método, se busca encontrar esos cinco conglomerados en agrupaciones hechas con otros métodos. Por ejemplo, se podría descubrir que el conglomerado número uno de está agrupación corresponde al conglomerado tres de una agrupación en ocho conglomerados hecha con AGNES, o al conglomerando cuatro de una agrupación en diez conglomerados hecha por k-means o al conglomerado dos de una agrupación en once conglomerados hecha por DIANA, o al conglomerado cinco de una agrupación en doce conglomerados hecha por PAM, es decir buscamos consistencia con todos los métodos, pero para hacer esto necesitamos saber cuán denso debe ser el conjunto de elementos que está alrededor de un centro, para decidir si sus elementos pertenecen a ese conglomerado, o cuán cerca deben estár los centros de dos conglomerados para concluir que son el mismo conglomerado. Las simulaciones hechas en el Capítulo 7 mostraron que los centros deben estár separados a una distancia mayor que aproximadamente tres veces la desviación estándar de los datos que están alrededor de los centroides para que todos los métodos-medida identifiquen correctamente los conglomerados.

Los datos reales constan de 244 variables compuestos por las respuestas de las pruebas prácticas en Internet, las respuestas a las pruebas reales en Internet, el número de veces que se tomo la prueba práctica, las respuestas a los cuestionarios, cuatro exámenes parciales en el aula de clase y el examen final. Se aplicaron todos los métodos de agrupación sobre el conjunto de datos reales, con particiones que van desde $k = 2…17$. De estás particiones se obtuvieron ocho conglomerados consistentes en todos los métodos. La descripción de las variables puede verse en el apéndice B.

Las siguientes tablas ilustran el proceso de búsqueda de consistencia descrito en el párrafo anterior.

## 8.2 Conglomerados Consistentes encontrados

La tabla 9 muestra las distancias de los conglomerados propuestos con los agrupamientos 8 al 13 producidos por el algoritmo PAM.

**Tabla 8.** Distancias entre los conglomerados propuestos y agrupaciones hechas con PAM

| | pam8 | pam9 | pam10 | pam11 | pam12 | pam13 |
|---|---|---|---|---|---|---|
| 1 | 4.098007* | 5.871373 | 4.790754 | 5.360215 | 7.63356 | 5.360215 |
| 2 | 5.533199* | 5.533199* | 6.017448* | 4.230319* | 8.142464* | 8.218582* |
| 3 | 4.310982* | 4.310982* | 4.692906* | 4.310982* | 2.290321* | 2.290321* |
| 4 | 0.00000* | 2.088017* | 1.617882* | 1.756282* | 1.756282* | 1.756282* |
| 5 | 6.202388* | 6.202388* | 5.31561* | 9.155972* | 8.624546* | 8.624546* |
| 6 | 5.934685* | 6.68987* | 8.095907* | 7.501467 | 7.501467 | 7.501467* |
| 7 | 6.686745 | 6.278480 | 7.35955* | 6.889798* | 6.889798* | 6.889798 |
| 8 | 7.465006* | 7.956188* | 4.046964* | 6.803137* | 5.137112* | 5.137112* |

Por ejemplo la primera la componente (1, pam8) =4.098007 *  indica que entre los ocho conglomerados agrupados por PAM, el más cercano al conglomerado uno tuvo una distancia de 4.098007 y el asterisco significa que fue significativamente menor (aproximadamente dos veces menor) que las distancias del conglomerado uno a los restantes siete conglomerados agrupados con PAM; en otras palabras, el conglomerado uno es equivalente a  uno de los ocho conglomerados agrupados con PAM.  El mismo procedimiento se utilizó  con cada uno de los métodos de aglomeración.

Las tabla 10  y 11 muestran las distancias de cada uno de los ocho conglomerados con agrupaciones de 8 a 13 utilizando k-means y AGNES.

**Tabla 9.** Distancias entre los conglomerados propuestos y agrupaciones hechas con kmeans

| kmeans8 | kmeans9 | kmeans10 | kmeans11 | kmeans12 | kmeans13 |
|---|---|---|---|---|---|
| 1 5.266627* | 4.770795* | 3.783970* | 8.699454 | 2.852489* | 3.448657* |
| 2 8.327217* | 7.743272* | 5.88094* | 13.1077 | 3 7.928221* | 8.121254* |
| 3 2.591567* | 3.70075* | 1.976184* | 7.688977* | 5.384847* | 10.49114* |
| 4 6.005535 | 5.193114 | 9.872636* | 4.31596* | 2.810208* | 3.735119* |
| 5 5.287364* | 5.287364* | 11.39032* | 7.56447* | 5.31561* | 8.059495* |
| 6 7.722075* | 7.609046* | 8.731228* | 8.04119* | 7.236067 | 7.028479 |
| 7 4.070429* | 6.715409 | 4.376462 * | 6.733958* | 6.296851 | 9.060917 |
| 8 7.248176* | 6.497458 * | 8.360609* | 3.745854* | 5.295798* | 10.300217 |

**Tabla 10.** Distancias entre los conglomerados propuestos y agrupaciones hechas con AGNES.

| | Anges8 | Anges9 | Anges10 | Anges11 | Anges12 | Anges13 |
|---|---|---|---|---|---|---|
| 1 | 1.912165* | 1.912165* | 1.912165* | 1.912165* | 1.912165* | 1.912165* |
| 2 | 3.870504* | 3.870504* | 9.979481* | 9.979481* | 9.979481* | 9.472282* |
| 3 | 3.144742* | 3.144742* | 3.14474* | 3.14474* | 3.14474* | 3.824975* |
| 4 | 1.782613* | 1.782613* | 1.78261* | 1.782613* | 1.782613* | 1.782613* |
| 5 | 11.358122* | 11.358122* | 3.366608* | 3.366608* | 3.366608* | 7.593979* |
| 6 | 9.165928* | 9.165928* | 9.165928 * | 9.165928 * | 9.165928 * | 9.165928* |
| 7 | 4.483380* | 4.483380* | 4.483380* | 4.483380* | 4.483380* | 4.483380*8 |
| 8.010943* | | 8.010943 * | 8.010943* | 8.010943* | 8.010943* | 8.010943* |

La tabla 12 muestra las distancias de los conglomerados propuestos con los agrupamientos 8 al 13 producidos por el algoritmo DIANA.

**Tabla 11.** Distancias entre los conglomerados propuestos y agrupaciones hechas con DIANA.

| | Diana8 | Diana9 | Diana10 | Diana11 | Diana12 | Diana13 |
|---|---|---|---|---|---|---|
| 1 | 6.986573 | 6.986573 | 6.986573 | 6.986573 | 6.065115 | 6.065115 |
| 2 | 12.644174 | 12.644174 | 12.045515 | 12.045515 | 12.045515 | 12.045515 |
| 3 | 5.323796* | 5.323796* | 5.323796* | 1.976184 * | 1.976184* | 3.824975* |
| 4 | 1.724289* | 1.724289* | 1.724289* | 1.724289* | 1.724289* | 1.724289* |
| 5 | 5.042444* | 5.042444* | 2.773175* | 2.773175* | 2.773175* | 2.773175* |
| 6 | 3.803685* | 4.133808* | 4.133808* | 4.133808* | 4.133808* | 4.133808* |
| 7 | 4.328247* | 4.328247* | 4.32824* | 4.328247* | 4.328247* | 4.328247* |
| 8 | 8.093568* | 8.093568* | 8.093568 | 8.093568 | 8.093568* | 8.093568 |

Si la distancia de un conglomerado propuesto al más conglomerado cercano de una de las agrupaciones, (de 8 a 13) es menor que cinco y si además esta distancia es

significativamente menor que la distancia del conglomerado propuesto a los restantes conglomerados (posee un asterisco), entonces los dos conglomerados son equivalentes.

## 8.3 Descripción de perfiles de los conglomerados encontrados

Los siguientes son los perfiles que describen a los estudiantes que componen cada uno de los ocho conglomerados. El análisis de los perfiles es independiente de la metodología usada para identificar los conglomerados, es decir, se describen a partir de los conglomerados encontrados, independientemente de la forma como estos encontraron

***Conglomerado 1:***

Grado de Claridad en el conglomerado: *Alto*

Promedio de veces que toma Exámenes de Práctica: 6.6

Temas donde es 25% superior: Ninguno

Temas donde es 25% inferior: Problemas verbales

Promedio de exámenes Parciales: 63%

Nota en el Examen Final: 57%

Tiempo Dedicado al curso: Adecuado

Expectativas del curso: Satisfactorias

***Conglomerado 2:***

Grado de Claridad en el conglomerado: *Mediano*

Promedio de veces que toma Exámenes de Práctica: 6.6

Temas donde es 25% superior: Ninguno

Temas donde es 25% inferior: Verbales, Algebraicos, Numéricos

Promedio de exámenes Parciales: 40%

Nota en el Examen Final: 40%

Tiempo Dedicado al curso: Mucho

Expectativas del curso: Bajas

***Conglomerado 3:***

Grado de Claridad en el conglomerado: *Alto*

Promedio de veces que toma Exámenes de Práctica: 6.7

Temas donde es 25% superior: Numéricos

Temas donde es 25% inferior: Ninguno

Promedio de exámenes Parciales: 48%

Nota en el Examen Final: 54%

Tiempo Dedicado al curso: Poco

Expectativas del curso: Bajas

***Conglomerado 4:***

Grado de Claridad en el conglomerado: *Alto*

Promedio de veces que toma Exámenes de Práctica: 4.7

Temas donde es 25% superior: Algebraicos, Geométricos, Verbales

Temas donde es 25% inferior: Ninguno

Promedio de exámenes Parciales: 84%

Nota en el Examen Final: 81%

Tiempo Dedicado al curso: Mucho

Expectativas del curso: Altas

***Conglomerado 5:***

Grado de Claridad en el conglomerado: *Alto*

Promedio de veces que toma Exámenes de Práctica: 5.8

Temas donde es 25% superior: Ninguno

Temas donde es 25% inferior: Verbales, Algebraicos, Geométricos

Promedio de exámenes Parciales: 44%

Nota en el Examen Final: 54%

Tiempo Dedicado al curso: Mucho

Expectativas del curso: Satisfactorias

***Conglomerado 6:***

Grado de Claridad en el conglomerado: *Mediano*

Promedio de veces que toma Exámenes de Práctica: 5.9

Temas donde es 25% superior: Ningunos

Temas donde es 25% inferior: Geométricos

Promedio de exámenes Parciales: 58%

Nota en el Examen Final: 43%

Tiempo Dedicado al curso: Poco

Expectativas del curso: Bajas

***Conglomerado 7:***

Grado de Claridad en el conglomerado: *Mediano*

Promedio de veces que toma Exámenes de Práctica: 4.8

Temas donde es 25% superior: Verbales, Algebraicos, Geométricos, Numéricos

Temas donde es 25% inferior: Ninguno

Promedio de exámenes Parciales: 81%

Nota en el Examen Final: 78%

Tiempo Dedicado al curso: Adecuado

Expectativas del curso: Altas

***Conglomerado 8:***

Grado de Claridad en el conglomerado: *Mediano*

Promedio de veces que toma Exámenes de Práctica:

Temas donde es 25% superior: Geométricos

Temas donde es 25% inferior: Ninguno

Promedio de exámenes Parciales: 61%

Nota en el Examen Final: 67%

Tiempo Dedicado al curso: Adecuado

Expectativas del curso: Satisfactorias

Los anteriores perfiles se identificaron sólo numéricamente. Aunque no se llego a definir con más detalle los perfiles asociados con los conglomerados, Se puede observar que cada conglomerado contiene aspectos únicos con respeto a fortalezas, debilidades, desempeño, y actitud. Para llegar a definir más profundamente los perfiles, se necesitaría más análisis desde el punto de vista de un experto en educación matemática.

# CAPÍTULO 9: CONCLUSIONES, TRABAJOS FUTUROS Y RELEVANCIA

## 9.1 Sobre la primera etapa: Manejo de Datos

El modelo de programación orientado a objetos se ajusta a las necesidades de este tipo de proyectos, pues es fácil, modificar agregar y remover conjuntos de datos, en este caso, semanas, cursos, estudiantes.

Se encontró que el paquete estadístico R es ideal para el desarrollo de este tipo de proyectos. Es de código abierto (*open source*) y posee un modelo robusto de programación que soporta todas las estructuras de datos conocidas. Además puede llamar código en C, C++ y Fortran. En este proyecto desarrollamos el código necesario para que algoritmo PAM pudiera trabajar con los índices de la librería cclust lo que nos llevo a uno de los mejores indicadores PAM/DB y PAM/Calinski

## 9.2 Sobre la segunda etapa: Capacidad de los métodos en la identificación conglomerados

Se encontró que con centros separados a dos unidades de distancia, y con una desviación estándar de 0.5 casi todos los métodos de validación identificaron los

conglomerados agrupados por todos los métodos de agrupación, excepto los agrupados por k-means.

Se puede ver a través de las simulaciones que con centros separados a dos unidades de distancia, se perdió la capacidad de identificar los conglomerados en las agrupaciones hechas por PAM a partir de una desviación estándar de 1.2.

Se determino a partir de las simulaciones que con centros separados a dos unidades de distancia, se perdió totalmente la capacidad de identificar los conglomerados en las agrupaciones hechas por k-means, a partir de una desviación estándar 0.74

Se concluye a partir de las simulaciones que con centros separados a dos unidades de distancia, se perdió totalmente la capacidad de identificar los conglomerados en las agrupaciones hechas por AGNES a partir de una desviación estándar 0.76

Se encontró que con centros separados a dos unidades de distancia, se perdió totalmente la capacidad de identificar los conglomerados en las agrupaciones hechas por DIANA a partir de una desviación estándar 0.74.

En general, se concluye que la distancia entre centros tiene que ser mayor que aproximadamente tres veces la desviación estándar de los datos para que todos los conglomerados puedan ser identificados por todas la medidas de validación.

## 9.3 Sobre la tercera etapa: los conglomerados obtenidos

Se encontraron ocho conglomerados consistentes. Estos se manifestaron en todas las agrupaciones hechas con todos los métodos.

Se encontró que cuatro de ellos son claros y bien definidos, es decir, bien separados uno del otro y se manifestaron en los cuatro métodos de agrupación. y con particiones que van desde ocho hasta trece.

Se encontraron cuatro conglomerados medianamente claros, es decir, aunque se manifestaron en las agrupaciones hechas por todos los métodos, hubo ciertas agrupaciones en donde la distancia del conglomerado propuesto al más cercano conglomerado de la partición no fue significativamente más pequeña que las distancias a los conglomerados restantes.

En general se considera que un conglomerado es claro si se manifiesta claramente en todas particiones hechas por todos los métodos y está bien separado de los demás. Además se considera que si un conglomerado es claro los centros deben estar separados a una distancia mayor que aproximadamente tres veces la desviación estándar de los puntos que están alrededor su centroide.

## 9.4 Trabajo futuro

**Desarrollo de una métrica**: Se debe desarrollar una métrica que asigne de manera equitativa los pesos y no favorezca tanto a magnitudes grandes como la métrica euclidiana, además mediante simulaciones se puede explorar los resultados de agrupar utilizando diferentes métricas.

**Imputación de datos faltantes:** Mas del 60% de los datos se perdieron, se inicio con una población de aproximadamente 400 estudiantes y se termino con una población de 106, por esta razón se considera que en futuros trabajos se deberían considerar técnicas de imputación datos.

**Trabajo necesario para asociar los conglomerados con perfiles educativos:** Para determinar la naturaleza precisa de estos perfiles es necesario recolectar más información, realizar entrevistas que tiendan a aclarar los perfiles que fueron medianamente claros. Además, las preguntas de los cuestionarios deben estar más orientadas a determinar los niveles de aprendizaje de los estudiantes

## 9.5 Relevancia

Éste proyecto fue realizado con datos educativos, para tratar de determinar los perfiles de los estudiantes que participan en el proyecto QUIZ del departamento de

Matemáticas del RUM, pero aunque está fue la población objetivo, la naturaleza del problema (más columnas que filas) es similar a problemas como el de expresión genética en experimentos con microarreglos y tratamiento de imágenes. Por esta razón se considera que esta metodología puede ser usada en la solución de esta clase de problemas.

## REFERENCIAS


[1] Berna. Frank, "Study Practices and Attitudes Related to Academic Success In a Distance Learning Programmed", Australia, 1993.
http://www.uned.ac.cr/servicios/global/administracion/costos/articulos/practicas.html

[2] Dudoit. Sandrine y Gentleman. Robert, "Cluster Analysis in DNA Microarray Experiments", Bioconductor Short Course Winter, California, 2002.
http://www.bioconductor.org/workshops/ShortCourse012302/lectures/lect5.pdf

[3] Romesburg. Charles, " Cluster Analysis for Research", Lifetime Learning Publication, Belmont, 1984.

[4] Behrouz. Minaei, "Using Genetic Algorithms for Data Mining Optimization in an Educational Web-based System". Michigan, 2003
http://www.ucm.es/BUCM/cee/doc/9902/9902.htm

[5] Joyanes A. Luis, "Programación Orientada a Objetos". McGraw-Hill, Madrid, 1998.

[6] Hartigan. John, "Clustering Algorithms", John Wiley & Sons, Nueva York, 1975.

[7] J. Daxin, "Cluster Analysis for Gene Expression Data: A Survey", Nueva York, 2002.
http://www.cse.buffalo.edu/~djiang3/publications/survey.pdf

[8] Gondar. José, "Análisis de Conglomerados", Madrid 2001.
http://www.estadistico.com/arts.html?20010723#subcap2

[9] Acuña. Edgar, "Notas del curso Temas en Estadística", Puerto Rico, 2003
http://math.uprm.edu/~edgar/esma683503.html

[10] Kaufman. Leonard, y Rousseeuw. Peter, " Finding Groups in Data An Introduction to Cluster Analysis". John Wiley &Sons. NuevaYork, 1990.

[11] D. Barbara, “An Introduction to Cluster Analysis for Data Mining”,California, 2000 http://www.ise.gmu.edu/~dbarbara/755/csurvey.pdf.

[12] Pollard. Katherine y Van der Laan. Mark “A Method to Identify Signicant Clusters in Gene Expression Data”, California, 2002. http://www.bepress.com/cgi/viewcontent.cgi?article=1002&context=ucbbiostat

[13] Weingessel, Adres y Dimitriadou, Eugenia, “An Examination Of Indexes For Determining The Numer of Cluster in Binary Data Sets”, Viena, 1999. http://www.wu-wien.ac.at/am/wp99.htm#29

# APÉNDICE A: CUESTIONARIOS

La siguiente cantidad es la que mejor representa el tiempo que estudio para mi clase de Cálculo (incluyendo las pruebas cortas por el Internet) cada semana:

1. menos de 3 horas
2. 3-6 horas
3. 6-9 horas
4. 9-12 horas
5. 12-15 horas
6. Más de 15 horas

La siguiente cantidad es la que mejor representa el tiempo que dedico a los materiales del Internet cada semana:

1. menos de 3 horas
2. 3-6 horas
3. 6-9 horas
4. 9-12 horas
5. 12-15 horas
6. Más de 15 horas

Si yo estuviese en una sección de Cálculo que no usa los materiales del Internet, la siguiente cantidad es la que mejor representa el tiempo que estudiaría para la clase de Cálculo cada semana;

1. menos de 3 horas
2. 3-6 horas
3. 6-9 horas
4. 9-12 horas
5. 12-15 horas
6. Más de 15 horas

Mi expectativa es recibir la siguiente nota:

1. A
2. B
3. C
4. D
5. F
6. W

Estudiar para las pruebas cortas de la página de Internet me ayudo a entender mejor el material de la clase:

1. Estoy totalmente de acuerdo (la aseveración es cierta la mayor parte de las veces)
2. Estoy parcialmente de acuerdo (la aseveración es cierta con suficiente frecuencia
3. Estoy parcialmente en desacuerdo (la aseveración no es cierta con suficiente frecuencia

4. Estoy en total desacuerdo (la aseveración no es cierta la mayor parte de las veces)

Mayormente el contenido de las pruebas cortas:

1. Era de material que no conocía

2. Era de material que había visto pero que no récordaba bien

3. Era de material que conocía y récordaba bien

El contenido de las pruebas cortas:

1. Casi siempre me ayudo a entender el material que se presento en el salón de clases

2. A veces me ayudo a entender el material que se presento en el salón de clases

3. Nunca o casi nunca me ayudo a entender mejor el material que se presento en el salón de clases

El contenido de las pruebas cortas:

1. Casi siempre me ayudo a entender el material que se presento en mis clases anteriores

2. A veces me ayudo a entender el material que se presento en mis clases anteriores

3. Nunca o casi nunca me ayudo a entender el material que se presento en mis clases anteriores

Tiene una computadora disponible en casa:

1. Si

2. No

Su padre o madre tiene un grado universitario:

1. Si

2. No

Que dificultades has tenido para tomar las pruebas por Internet?  Incluye a continuación cualquier comentario que estimes pertinente acerca de las pruebas de Internet.

# APÉNDICE B: DESCRIPCIÓN DE LAS VARIABLES

Primera semana
Variables 1-6: Primera prueba de práctica
Variables 7-12: Última prueba de práctica
Variables 13-18: Promedio de todas las pruebas de práctica
Variable 19: Número de veces que tomo la prueba de práctica
Variables 20-25: Prueba real

Segunda semana
Variables 26-33: Primera prueba de práctica
Variables 34-41: Última prueba de práctica
Variables 42-49: Promedio de todas las pruebas de práctica
Variable 50: Número de veces que tomo la prueba de práctica
Variables 51-58: Prueba real

Tercera semana
Variables 59-63: Primera prueba de práctica
Variables 64-68: Última prueba de práctica
Variables 69-73: Promedio de todas las pruebas de práctica
Variable 74: Número de veces que tomo la prueba de práctica
Variable 75-79: Prueba real

Cuarta semana
Variables 80-84: Primera prueba de práctica
Variables 85:-89: Última prueba de práctica
Variables 90-94: Promedio de todas las pruebas de práctica
Variable 95: Número de veces que tomo la prueba de práctica
Variable 96-100: Prueba real

Quinta semana
Variables 101-106: Primera prueba de práctica
Variables 107:-112: Última prueba de práctica
Variables 113:-118: Promedio de todas las pruebas de práctica
Variable 119: Número de veces que tomo la prueba de práctica
Variables 120-125: Prueba real

Sexta semana
Variables 126-131: Primera prueba de práctica

Variables 132-137: Última prueba de práctica
Variables 138-143: Promedio de todas las pruebas de práctica
Variable 144: Número de veces que tomo la prueba de práctica
Variables 145-150: Prueba real

Séptima semana
Variables 151-156: Primera prueba de práctica
Variables 157-162: Última prueba de práctica
Variables 163-168: Promedio de todas las pruebas de práctica
Variable 169: Número de veces que tomo la prueba de práctica
Variables 170-175: Prueba real

Octava semana
Variables 176-182: Primera prueba de práctica
Variables 183-189: Última prueba de práctica
Variables 190-196: Promedio de todas las pruebas de práctica
Variable 197: Número de veces que tomo la prueba de práctica
Variable 198-204: Prueba real

Novena semana
Variables 205-210: Primera prueba de práctica
Variables 211:216: Última prueba de práctica
Variables 217-222: Promedio de todas las pruebas de práctica
Variable 223: Número de veces que tomo la prueba de práctica
Variables 224 -229: Prueba real

Cuestionarios: Variables 230-239
Cuatro exámenes en el salón clase 240-243
Examen final en el salón clase: 244

Pesos
Cada una de las preguntas de la prueba de práctica: 1
Cada una de las preguntas de la prueba real: 1
Número de veces que se tomo la prueba de práctica: 1
Cada una de las preguntas de los cuestionarios: 2
Cada uno de los exámenes parciales: 50
Examen Final: 50