**DISEÑO EXPERIMENTO:**

**Fase 1:***Entiende y delimita el problema sobre el que se va a hacer el experimento. Identifica la información más relevante del problema del experimento.*

El método Clustering K Means se utiliza para agrupar información respecto a un factor de similitud deseado. Por ende, se quiere observar como varía el comportamiento (temporal y correctitud) del algoritmo implementado con respecto a diferentes factores variables.

En el problema en cuestión, el algoritmo de clustering K Means fue implementado con el objetivo de realizar agrupaciones sobre los Itemsets frecuentes, donde para llevar a cabo esta labor, es necesario ingresar la cantidad agrupaciones o clusters deseados, por tal, este será uno de los factores que irá variando a lo largo del experimento, posteriormente se establecerán los niveles para este experimento

Asimismo, la cantidad de itemsets frecuentes que se van a agrupar afectará la velocidad de procesamiento del algoritmo y tal vez el grado de similitud de los itemsets dentro de cada cluster. Por lo que este será otro factor variable que se va a estar alterando. ¿Cómo se comporta con 20 itemsets, con 40, con 60? Etc…

Lo que se desea finalmente encontrar es cuál es el nivel más óptimo para la generación de clusters. Considerando no solo la rapidez de ejecución si no también la correctitud (similitud) de los agrupamientos.

**Fase 2:**

*Elige las variables de respuesta a medir, teniendo en cuenta que son aquellas variables que mejor caractericen el problema.*

Variables de respuesta:

* Tiempo que tarda el algoritmo en construir los clústeres. Este tiempo irá cambiando según la cantidad de ítems que estén registrados y el número de clústeres que se le soliciten construir. Debido a que los tiempos de ejecución tienden a ser muy rápidos, este se medirá en milisegundos con la clase Stopwatch de Visual Studio.

* Correctitud de los agrupamientos. Medición de la similitud de los itemsets frecuentes dentro de cada clúster. Se va a considerar la desviación estándar de los atributos de los itemsets dentro de cada clúster como unidad representativa para medir la similitud de los elementos dentro de cada agrupamientos.  
  Se escoge la desviación estándar como la herramienta para evaluar la correctitud ya que esta es una medición de la dispersión para variables de razón (cuantitativas). De esta forma se puede observar que tan “disperso” está cada clúster.

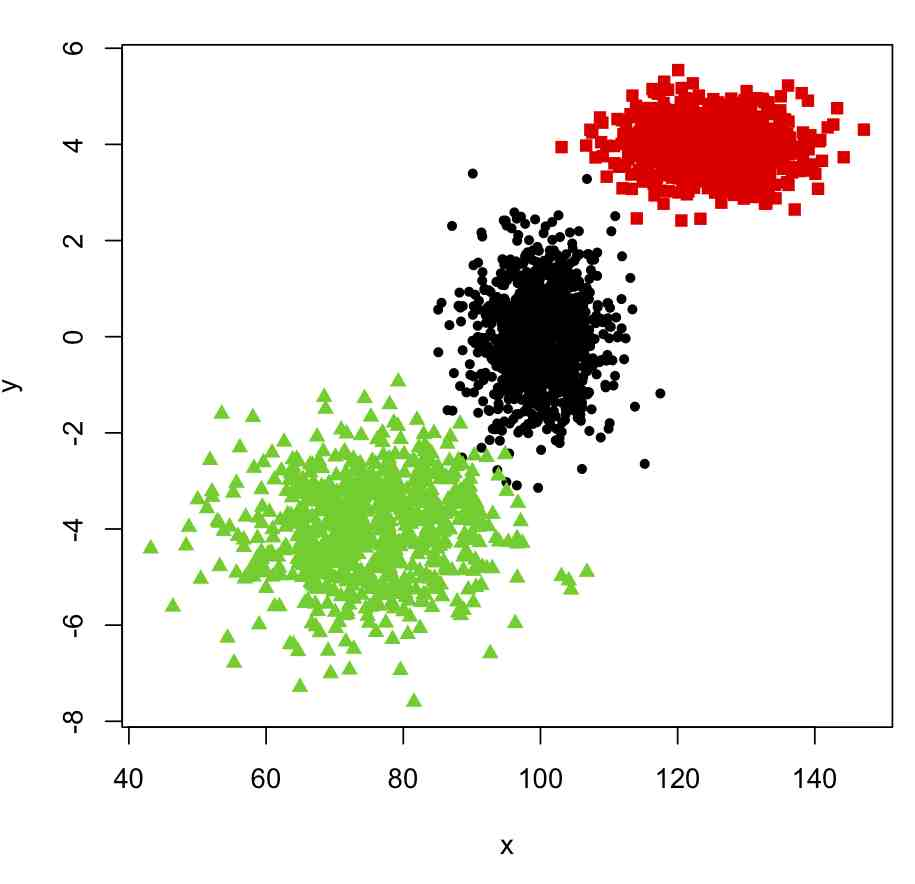
Para ilustrar mejor esta idea, se abordará el siguiente ejemplo:

Sea *n* un tratamiento del experimento. Cada itemset dentro de cada cluster *ci* de *n* tiene dos propiedades que lo definen (*x* e *y*).

El primer paso consiste en calcular la desviación estándar de cada uno de los clusters *ci*.

El segundo paso consiste en tomar cada una de esas desviaciones y calcular una desviación estándar general (la desviación de las desviaciones), para cada propiedad, de *n*. Esos serán los valores que definan la correctitud de *n*.

Explicación gráfica:



Primer paso: Calcular la desviación estándar de las dos propiedades x e y en cada cluster C1, C2, C3. Se obtendrían entonces Sx1, Sx2 y Sx3, Sy1, Sy2 y Sy3.

Segundo paso: Calcular la desviación estándar de las desviaciones estándares de cada clúster,. Es decir: SxC es igual a la desviación de los valores de Sx1, Sx2 y Sx3. Se realiza el mismo procedimiento para SyC.

Los valores SxC y SyC definen la similitud del clustering realizado por el algoritmo.

.

**Fase 3:**

*Determina los factores de estudio, haciendo claridad que son estos factores los que afectan los resultados del experimento.*

Factores de estudio:

* Cantidad de agrupamientos.
* Cantidad de itemsets frecuentes.

Factores controlables:

* Cantidad de clusters demandados al algoritmo.
* Cantidad de itemsets frecuentes.
* Cantidad de clientes que se cargan.
* Cantidad de ítems que se cargan.
* Cantidad de transacciones que se cargan.

Factores no controlables.

* Ejecución de procesos en segundo plano.
* Fragmentación de la unidad de almacenamiento.

Unidades experimentales:

* Número de Clusters solicitados al algoritmo.
* Número de itemsets frecuentes que el algoritmo deberá agrupar.

**Fase 4:**

*Selecciona el diseño experimental adecuado teniendo en cuenta los niveles de cada factor*

La siguiente tabla presenta los niveles para la cantidad de itemsets frecuentes que serán utilizados en el experimento.

|  |  |
| --- | --- |
| Cantidad de itemsets frecuentes (n) | Nivel de itemsets frecuentes |
| 20 | 1 |
| 40 | 2 |
| 60 | 3 |
| 80 | 4 |
| 100 | 5 |

La siguiente tabla presenta los niveles para la cantidad de clústeres que serán utilizados en el experimento.

|  |  |
| --- | --- |
| Cantidad de clústeres | Nivel de clusters |
| 1 | 1 |
| n\*1/4 | 2 |
| n\*1/2 | 3 |
| n\*3/4 | 4 |
| n | 5 |

Las multiplicaciones de la cantidad de clusters se explican así:  
Como no es lo mismo agrupar por ejemplo 10 elementos en 5 grupos que juntar 100 elementos en 5 grupos (en cuestiones de similitud de los elementos en los grupos), los niveles del número de clusters se tomaron como multiplicaciones por una proporción. De forma que, el nivel 2 para 20 itemsets, sería equivalente a construir 20\*0,25=>5 clusters. Mientras que, para 100 itemsets, este nivel construiría 100\*0,25=>25 clusters.

**Fase 5:**

*Define una planeación para el desarrollo del trabajo experimental*

Ahora bien, cada tratamiento se ejecutará 1000 veces y se registrarán sus resultados, para posteriormente realizar una ANOVA (*Analysis of variance)* y encontrar el tratamiento más óptimo, dependiendo de la calidad (similitud) del resultado y el tiempo de desarrollo para el mismo.

A continuación, se presentan los tratamientos que serán utilizados en este experimento.

|  |  |  |
| --- | --- | --- |
| Nivel de ítems frecuentes | Nivel de clusters | Tratamiento |
| 1 | 1 | 1 |
| 2 | 2 |
| 3 | 3 |
| 4 | 4 |
| 5 | 5 |
| 2 | 1 | 6 |
| 2 | 7 |
| 3 | 8 |
| 4 | 9 |
| 5 | 10 |
| 3 | 1 | 11 |
| 2 | 12 |
| 3 | 13 |
| 4 | 14 |
| 5 | 15 |
| 4 | 1 | 16 |
| 2 | 17 |
| 3 | 18 |
| 4 | 19 |
| 5 | 20 |
| 5 | 1 | 21 |
| 2 | 22 |
| 3 | 23 |
| 4 | 24 |
| 5 | 25 |