
UNIVERSIDAD
ANDRES BELLO

Universidad Andrés Bello
Facultad de Ingeniería

Ingeniería Civil Informática

# TAREA4

# APRENDIZAJE NO-SUPERVISADO

## SISTEMAS INTELIGENTES

DAVID ANDRÉS MOLINA GARRIDO

Profesor Alejandro Figueroa
Ayudante Alexander Espina

*Santiago - Chile.*

*Octubre, 2016.*

# Introducción

El siguiente informe consta de aprender los conceptos relacionados con aprendizaje no-supervisado tanto los modelos, métricas de evaluación, y la metodología de trabajo. Para esto, utilizaremos dos métodos de agrupamiento de datos, K-MEANS y DBSCAN.

El set de datos que se entrega a los algoritmos de agrupamiento (clustering) se obtiene de un conjunto de preguntas provenientes de un sitio web (cQA) de Yahoo!, llamado YahooAnswer, las cuales fueron etiquetadas según las siguientes clases:

-**Opiniones** (Preguntan por opiniones sobre un tema).
-**No Respondibles** (La respuesta no puede ser encontrada en ninguna página web).
-**Particulares** (Búsqueda de la aplicación del conocimiento a una situación particular).
-**Archivadas** (Preguntas hechas con anterioridad en YahooAnswer, con respuesta válida).
-**Web** (Tiene respuestas validas en la web).
-**Complejas** (La respuesta es poco probable que esté en solo un sitio web, por su complejidad).
-**Otros** (No cae en ninguna clase anterior).

Posteriormente, estas preguntas son representadas por un vector utilizando las palabras contenidas en cada pregunta como los atributos de este vector (como vimos en la Tarea2).

A diferencia de las Tareas anteriores nuestros datos no tendrán como una variable dependiente la etiqueta que anteriormente le dimos a cada una de nuestras instancias, sino que todas las instancias tendrán el mismo status.

Lo que se intenta obtener con K-MEANS es encontrar grupos de observaciones (*clusters*) con características semejantes y éstos ser diferentes a los otros grupos o *clusters*. Técnicamente se intenta maximizar la variación inter-*cluster* y minimizar la variación intra-*cluster*.

Para el caso de DBSCAN (*Density-based spatial clustering of applications with noise*), lo que se intenta encontrar son grupos diferenciados por medio de la densidad de éstos, así como también ruido, datos que no se asemejan a ningún grupo. Para esto, DBSCAN ocupa dos parámetros: un radio $\varepsilon$ , para poder revisar la vecindad a un punto específico; y una cantidad de puntos mínimos, que depende del radio $\varepsilon$.

# Descripción del Problema

Primeramente, para generar la salida de K-MEANS utilizaremos los datos representados vectorialmente en la Tarea 2, éstos se utilizan para generar los *clusters* que, a su vez, estarán determinados por 8 centroides que representan nuestras 8 clases. Se utilizará la implementación de *Clustering* K-MEANS de Yakmo y el software de minería de datos WEKA.

Los centroides son la media de todos los puntos de cada *cluster*. Al comienzo, se elegirán k centroides que se irán recalculando, en nuestro caso k=8 , ya que el objetivo es que cada *cluster* represente una clase. En el siguiente ejemplo con k=2 vemos el modo de reasignación de los centroide en sus *clusters*.

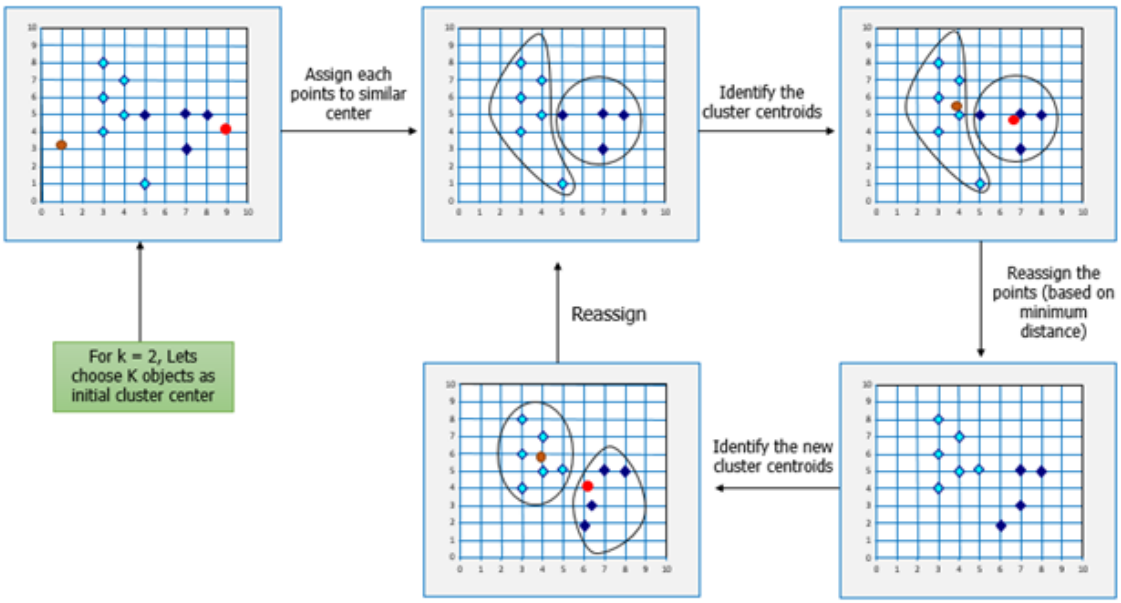


Imagen1: Ejemplo reasignación de un centroide.

La reasignación se basa en las medias de todos los puntos de cada *cluster*, La distancia desde un punto (instancia) xj al centroide mj se calcula:

$$dist(x_i, m_j) = \sqrt{(x_{i1} - m_{j1})^2 + (x_{i2} - m_{j2})^2 + \ldots + (x_{ir} - m_{jr})^2}$$

Imagen2: Distancia de un punto al centroide.

Para generar la representación de agrupamiento en DBSCAN se utilizó el software de minería de datos WEKA, utilizando una función de distancia euclidiana.

Se le entrega nuestro set de datos, para posteriormente designar un radio ε , con el cual poder revisar la vecindad a un punto específico; y una cantidad de puntos mínimos, que depende del radio ε, para poder incluirlos dentro de un *cluster*.

Este método recursivo se representa en la siguiente imagen:

En el ejemplo ε = 1 y los puntos mínimos de la vecindad = 4 .

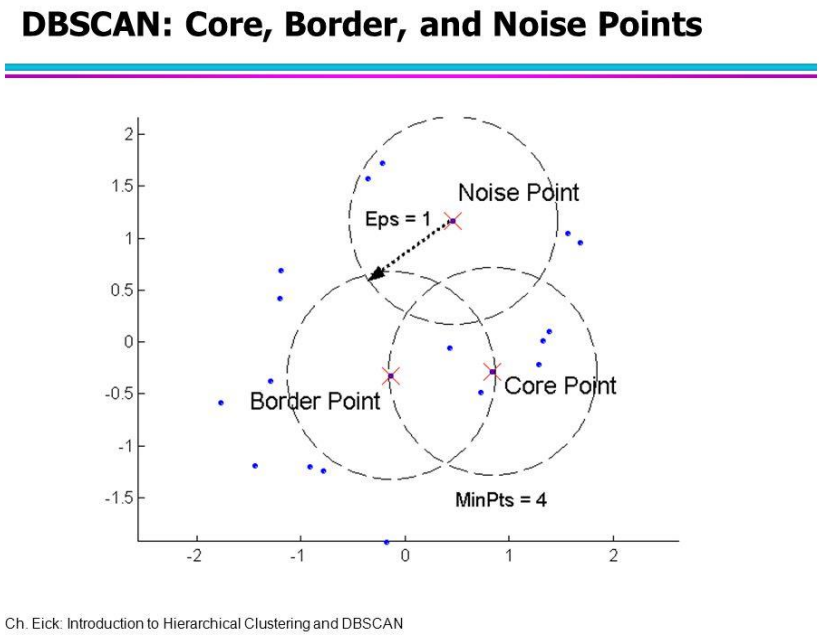


Imagen3: Variables DBSCAN .

**Nota:** DBSCAN no funcionó como debería, me arrojaba datos sin sentidos con una implementación y con la implementación en WEKA me arrojaba error, se intentó buscar la solución, sin éxito.

# Análisis del Resultado

Para que en las distintas clases de las preguntas sea viable su modelamiento, se le entregó un valor numérico que las represente, estos fueron:

| Clase | Número |
|---|---|
| Archivada | 1 |
| Archivadas y Web | 2 |
| Complejas | 3 |
| No-respondibles | 4 |
| Opiniones | 5 |
| Otros | 6 |
| Particulares | 7 |
| Web | 8 |

Imagen4: Tabla de clase y su número asignado.

A través de la implementación de *Clustering* K-MEANS de Yakmo se les entregó nuestro set de datos, para luego asignar la clase de su instancia a un *cluster* en particular.

Se asignó para cada *cluster* un valor dentro de las 8 clases, esto para que cada *cluster* se represente como una clase y sea posible el cálculo de métricas posteriormente. Entonces, a cada *cluster* se le asignó el valor más representativo, ósea, el que más apareció en ese *cluster*, los datos son los siguientes:

| **Cluster** | **Clase** |
|---|---|
| C1 | 1 |
| C4 | 2 |
| C3 | 3 |
| C5 | 4 |
| C6 | 5 |
| C0 | 6 |
| C2 | 7 |
| C7 | 8 |

Imagen5: Tabla de clase por cada *cluster.*

Como habíamos visto en la Tarea 1 la distribución de nuestras preguntas por clase (Imagen6 ) **Archivadas y Web**, **Particulares** y **Opiniones** fueron las mayores, es por esto que en la distribución por *cluster* también fueron las mayoritarias por cada uno de los 8 *clusters*. Por lo tanto, se procedió a desempatar, asignándole la segunda clase mayoritaria a los *clusters*, esto porque la etiqueta mayor ya ha sido usada anteriormente.

| Clases | Etiqueta | Total Preguntas | Probabilidad |
|---|---|---|---|
| Archivada | 1 | 40 | 0,06153846 |
| Archivadas y Web | 2 | 130 | 0,2 |
| Complejas | 3 | 13 | 0,02 |
| No-respondibles | 4 | 14 | 0,02153846 |
| Opiniones | 5 | 207 | 0,31846154 |
| Otros | 6 | 60 | 0,09230769 |
| Particulares | 7 | 134 | 0,20615385 |
| Web | 8 | 52 | 0,08 |

Imagen 6: Distribución de preguntas.

El detalle de cada *Cluster* según la clase y la cantidad es el siguiente:

| C0 | | Clase Mayoritaria |
|---|---|---|
| 1 | 1 | 6 |
| 2 | 4 | |
| 3 | 0 | |
| 4 | 0 | |
| 5 | 12 | |
| 6 | 1 | |
| 7 | 12 | |
| 8 | 0 | |

| C1 | Cantidad | Clase Mayoritaria |
|---|---|---|
| 1 | 0 | 1 |
| 2 | 1 | |
| 3 | 0 | |
| 4 | 0 | |
| 5 | 0 | |
| 6 | 1 | |
| 7 | 0 | |
| 8 | 0 | |

| C2 | Cantidad | Clase Mayoritaria |
|---|---|---|
| 1 | 6 | 7 |
| 2 | 29 | |
| 3 | 4 | |
| 4 | 1 | |
| 5 | 55 | |
| 6 | 6 | |
| 7 | 37 | |
| 8 | 8 | |

| C3 | Cantidad | Clase Mayoritaria |
|---|---|---|
| 1 | 0 | 3 |
| 2 | 0 | |
| 3 | 0 | |
| 4 | 0 | |
| 5 | 1 | |
| 6 | 0 | |
| 7 | 1 | |
| 8 | 0 | |

| C4 | | Clase Mayoritaria |
|---|---|---|
| 1 | 2 | 2 |
| 2 | 14 | |
| 3 | 0 | |
| 4 | 0 | |
| 5 | 21 | |
| 6 | 4 | |
| 7 | 15 | |
| 8 | 1 | |

| C5 | Cantidad | Clase Mayoritaria |
|---|---|---|
| 1 | 0 | 4 |
| 2 | 0 | |
| 3 | 0 | |
| 4 | 1 | |
| 5 | 1 | |
| 6 | 0 | |
| 7 | 2 | |
| 8 | 0 | |

| C6 | Cantidad | Clase Mayoritaria |
|---|---|---|
| 1 | 31 | 5 |
| 2 | 81 | |
| 3 | 9 | |
| 4 | 12 | |
| 5 | 116 | |
| 6 | 48 | |
| 7 | 67 | |
| 8 | 42 | |

| C7 | Cantidad | Clase Mayoritaria |
|---|---|---|
| 1 | 0 | 8 |
| 2 | 1 | |
| 3 | 0 | |
| 4 | 0 | |
| 5 | 1 | |
| 6 | 0 | |
| 7 | 0 | |
| 8 | 1 | |

Imagen7: Instancias por Clusters Yakmo.

Los 10 datos más representativos de todos los conjuntos de *clusters* fueron los siguientes:

En orden de más cercano a su centroide, con la etiqueta respectiva.

| Más cerca de su Centroide | Clase | Cluster donde pertenece |
|---|---|---|
| 1 | 5 | C6 |
| 2 | 5 | C6 |
| 3 | 5 | C6 |
| 4 | 2 | C6 |
| 5 | 5 | C2 |
| 6 | 7 | C2 |
| 7 | 7 | C2 |
| 8 | 5 | C6 |
| 9 | 2 | C4 |
| 10 | 5 | C6 |

Imagen8: 10 Instancias más representativas.

Como vemos, las 10 instancias más representativas de los *clusters*, estuvieron solo en 3 *clusters*: C2, C4 y C6. Esto se entiende ya que son los que tienen la mayor cantidad de datos con mayor cantidad de instancias. Por otro lado, se establece la relación que a excepción del 4to puesto, siempre son la etiqueta de la clase mayoritaria a su *cluster*, con lo que se deduce que esa etiqueta es la más representativa en cuanto a atributos en su *cluster*.

## Instancias por *Cluster*

Basándose en el Software de minería de datos WEKA, se obtuvieron los siguientes resultados, con relación a las instancias por *cluster* y otros:

El agrupamiento que se generó con 8 centroides (que representan las 8 clases) fue el siguiente:

| Clustered Instances | | |
|---|---|---|
| Cluster | nº Instancias | Porcentaje |
| 0 | 1 | 0% |
| 1 | 349 | 54% |
| 2 | 1 | 0% |
| 3 | 198 | 30% |
| 4 | 7 | 1% |
| 5 | 1 | 0% |
| 6 | 12 | 2% |
| 7 | 81 | 12% |
| | 650 | 99% |

Imagen9: Instancias por *Clusters* WEKA.

Estos datos se basaron usando una función de distancia euclidiana, y se recalcularon 7 veces los *clusters* hasta llegar al estado final.

Como vemos en la Imagen8 se generaron 3 *clusters* que cubrieron la mayoría de las instancias, que tiene relación con las 3 clases que más se etiquetaron en el set de datos. Esto quiere decir que encontró semejanzas en cada uno de estos 3 *cluster* y diferencia respecto a los otros 4, ósea, se maximizó la variación Inter-*Cluster* y minimizó la Intra-*Cluster*.

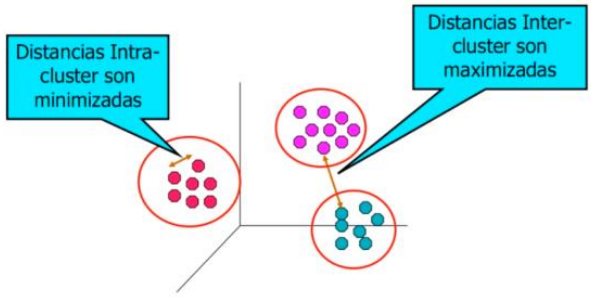


Imagen10: Distancias *Clusters*.

## Métricas

Nuestros datos con la implementación en Yamko, representados en una matriz de confusión:

| | C1 | C4 | C3 | C5 | C6 | C0 | C2 | C7 |
|---|---|---|---|---|---|---|---|---|
| | Prediccion / Clase | | | | | | | |
| Real | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| 1 | 0 | 2 | 0 | 0 | 31 | 1 | 6 | 0 |
| 2 | 1 | 14 | 0 | 0 | 81 | 4 | 29 | 1 |
| 3 | 0 | 0 | 0 | 0 | 9 | 0 | 4 | 0 |
| 4 | 0 | 0 | 0 | 1 | 12 | 0 | 1 | 0 |
| 5 | 0 | 21 | 1 | 1 | 116 | 12 | 55 | 1 |
| 6 | 1 | 4 | 0 | 0 | 48 | 1 | 6 | 0 |
| 7 | 0 | 15 | 1 | 2 | 67 | 12 | 37 | 0 |
| 8 | 0 | 1 | 0 | 0 | 42 | 0 | 8 | 1 |
| Total por Clústers | 2 | 57 | 2 | 4 | 406 | 30 | 146 | 3 |
| Prob por Clústers | 0,00307692 | 0,08769231 | 0,00307692 | 0,00615385 | 0,62461538 | 0,04615385 | 0,22461538 | 0,00461538 |
| Total | 650 | | | | | | | |

Imagen 11: Matriz de confusión de datos por *clusters*.

La matriz de confusión representa en las columnas nuestras clases mayoritarias pertenecientes a un *cluster*; y en las filas, la clase a la cual pertenece esa distribución de datos.

Como se muestra en la Imagen10 cada *clusters* no es fiel representante de su clase, existen más clases en un mismo *clusters*, además no siempre se representa la clase mayoritaria ya que ésta es mayoritaria de igual manera en otros *clusters*.
Es por esto, que existe un alto porcentaje de error en el agrupamiento que se genera, pero ese error está relacionado a nuestro etiquetado, es posible que según nuestros datos y atributos de las instancias el método K-MEANS lo haya particionado de buena manera.

Las métricas de cada *cluster* con su respectiva clase mayoritaria fueron:

| | | | |
|---|---|---|---|
| Clase 1 Cluster C1 | | | |
| Precision | 0 | Purity | 0 |
| Recall | 0 | Entropy | 0,008558252 |
| f1-score | #¡DIV/0! | | |
| | | | |
| | | | |
| Clase 2 Cluster C4 | | | |
| Precision | 0,10769231 | Purity | 0,245614035 |
| Recall | 0,24561404 | Entropy | 0,102641095 |
| f1-score | 0,14973262 | | |
| | | | |
| | | | |
| Clase 3 Cluster C3 | | | |
| Precision | 0 | Purity | 0 |
| Recall | 0 | Entropy | 0,008558252 |
| f1-score | #¡DIV/0! | | |
| | | | |
| | | | |
| Clase 4 Cluster C5 | | | |
| Precision | 0,07142857 | Purity | 0,25 |
| Recall | 0,25 | Entropy | 0,015065222 |
| f1-score | 0,11111111 | | |
| | | | |
| | | | |
| Clase 5 Cluster C6 | | | |
| Precision | 0,56038647 | Purity | 0,285714286 |
| Recall | 0,28571429 | Entropy | 0,141362952 |
| f1-score | 0,37846656 | | |
| | | | |
| | | | |
| Clase 6 Cluster C0 | | | |
| Precision | 0,01666667 | Purity | 0,033333333 |
| Recall | 0,03333333 | Entropy | 0,068267774 |
| f1-score | 0,02222222 | | |
| | | | |
| | | | |
| Clase 7 Cluster C2 | | | |
| Precision | 0,2761194 | Purity | 0,253424658 |
| Recall | 0,25342466 | Entropy | 0,161309137 |
| f1-score | 0,26428571 | | |
| | | | |
| Clase 8 Cluster C7 | | | |
| Precision | 0,01923077 | Purity | 0,333333333 |
| Recall | 0,33333333 | Entropy | 0,011937436 |
| f1-score | 0,03636364 | | |

Imagen 12: Métricas por *clusters*.

Como vemos, el *Cluster* 6 que representa a la clase 5 (Opiniones) tuvo los mayores índices, esto en relación a lo que logramos concluir de la Tarea 2:

1.Opiniones fue la clase con mayores instancias etiquetadas.
2.El sitio tiende a generar mayor probabilidad de preguntas en esta clase, generando opiniones de los demás usuarios.
3.No se encontraban las Categorías repartidas uniformemente por clase, y además las que existían se tendían a ir por la clase Opiniones.

Por otro lado, los *cluster* 1,3 y 7 son los con menores instancias, pero éstas en su mayoría no corresponden a su clase mayoritaria. Es por esto, su baja *Precision* y *Recall*, que corresponden a que tienen muchos falsos positivos y falsos negativos.

El *Purity* y Entropía total de los *clusters* fue el siguiente:

| | |
|---|---|
| Purity Total | 0,26153846 |
| Entropy Total | 0,51770012 |

Imagen 13: *Purity* y *Entropy* totales.

El purity total se obtuvo de la sumatoria de dividir el total de instancias de un *cluster* i por el total de datos (650), multiplicado por el *purity* del *cluster* i:

El *purity* total nos dice que existe poca pureza en ellos, que se corresponde y se entiende de manera específica en la *precision* y *recall*. Esta poca pureza se basa en que cada *cluster* no toma de manera regular su clase mayoritaria, sino que varía con otras clases.

Por otro lado, para la entropía se utilizó logaritmo en base 8 para cada *cluster*, con el propósito de hacerla más entendible. La entropía total nos dice que el grado de certeza de los *clusters* con respecto a su clase mayoritaria es “regular”, casi ½ de certeza en la totalidad de *clusters*.

# Conclusión

Como conclusión podemos ver que el método de agrupamiento en cluster nos indica de manera más visible la información contenida en nuestro set de datos y los agrupa de manera de encontrar semejanzas en ellos.

Como habíamos visto en K-MEANS la información resultante tanto de la distribución de instancias etiquetadas, el particionamiento de Yakmo y el que realizamos en WEKA correspondió y dio como resultado 3 cluster específicos, que se relacionan con:
-Opiniones
-Particulares
-Archivadas y Web

En la implementación con Yakmo, en particular el cluster 6, tuvo los mejores índices, además fue el cluster que mayor cantidad de datos agrupo, con un porcentaje de 0.6246, por lo que se vio que la tendencia fue muy fuerte en ese cluster, reuniendo características específicas y diferenciadas del resto.

Por otra parte, se deduce que los atributos como palabras quizá no es la mejor manera de representar vectorialmente las instancias, esto ya que utilizando métodos de agrupamiento no se reflejan de la mejor manera las diferencias entre una y otra.