

Minería de datos **Práctica 1:Clustering** **knn-means**

Jose Ignacio Sánchez
Josu Rodríguez

24 de octubre de 2014

ÍNDICE DE CONTENIDO

1. Introducción	1
2. Recursos	1
3. Clasificación NO-supervisada o <i>Clustering</i>	1
3.1. Clustering <i>k-means</i>	1
4. Diseño	1
4.1. Algoritmo en pseudocódigo	5
5. Implementación	6
5.1. Análisis del conjunto de datos para decidir la conveniencia de la normalización	6
5.2. Problemas encontrados y soluciones adoptadas	6
5.2.1. Problema con la normalización de los atributos	6
5.2.2. Problema de generación de excesivas instancias	6
5.2.3. Instancias repetidas en los gráficos	6
5.2.4. Ausencia de representación para el cluster 0	7
6. Validación del <i>software</i>	7
6.1. Diseño del banco de pruebas	7
7. Análisis de resultados	7
7.1. Modificando inicializaciones	7
7.2. Modificando distancia Minkowski	7
7.3. Criterios de convergencia	7
7.3.1. Número fijo de iteraciones	7
7.3.2. Disimilitud entre <i>codebooks</i>	7
7.4. Distintas métricas	7
7.4.1. Manhattan	7
7.4.2. Euclídea	7
7.4.3. Minkowski	7
8. Clasificación supervisada respecto de	7
9. Conclusiones	7
9.1. Técnicas de clustering: motivación	7
9.2. Conclusiones generales	7
9.3. Puntos débiles y propuestas de mejora	8
10. Valoración subjetiva	8

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

1.	Clusters:Separación y cohesión	2
2.	Coeficiente silhouette para un punto i	3
3.	Esquema de dependencias del sistema	4

1. Introducción

El objetivo principal de esta práctica es obtener la capacidad de formular un algoritmo de aprendizaje automático de clasificación **No-Supervisada**. Por otra parte, se trabajarán la capacidad de sintetizar una técnica de aprendizaje automático no-supervisado, conocer su coste computacional así como sus limitaciones de representación y de inteligibilidad

2. Recursos

- PC con aplicación Weka.
- Bibliografía.
- Librerías de Weka.
- Manual de Weka.
- Guía de la práctica.
- Ficheros para los datos de la práctica: [food.arff](#), [colon.arff](#).
- Otros ficheros que no están en formato *.arff*:
 - En formato *.txt*: [ClusterData.atributos.txt](#) (este fichero si tiene la clase asociada para evaluar la calidad del *clustering* en [ClusterData.clase.txt](#)).
 - En formato *.csv* [bank-data.csv](#)clustering

3. Clasificación NO-supervisada o *Clustering*

(Definición) 3.1 Técnicas de aprendizaje donde no hay un conocimiento a priori, donde agrupa las instancias sin atributos dependientes pre-especificados. Los algoritmos clustering son un método común de aprendizaje no supervisado.

3.1. Clustering *k-means*

4. Diseño

Estructuramos la ejecución del algoritmo en fases como se puede ver en la figura 3 , las cuales se detallan a continuación.

Primera fase: carga de datos y configuración

Inicialmente se carga la configuración establecida por el usuario en el fichero **kmeans.conf**, es decir: path del fichero y su formato, tipo de inicialización para el *codebook*, número de clusters, distancia a utilizar, número de clusters deseados, elección manual o automática sobre la normalización y diversas opciones más, especificando datos sobre el fichero que se utilizará para las instancias.

Segunda fase: Preproceso de datos

En el preproceso de datos se normalizará o no, dependiendo del parámetro indicado por el usuario. Si el parámetro es 0 no se normalizará, si es 1 se normaliza y si es 2 se hará uso del método experimental para decidir la idoneidad de la normalización.

Tercera Fase: Algoritmo K-means

En esta fase se implementa el algoritmo **K-means**.

1. En primer lugar inicia los *centroides* con el criterio establecido por el usuario, o la matriz de bits de pertenencias.
2. Recorre las instancias del conjunto y calcula la distancia a cada uno de los *codeword* actualizando la matriz de bits de pertenencia, el valor del bit es uno si es el centroide más cercano a la instancia.
3. Se calcula de nuevo el vector promedio para cada cluster.
4. Iterar los pasos dos y tres hasta converger.

Cuarta Fase: Evaluación

Existen distintos métodos propuestos de evaluación de Clustering, los que conocemos actualmente pueden dividirse en dos grupos:

- *Extrinsic methods*: Se aplican cuando disponemos de la etiqueta de las instancias, asignando una puntuación en función de las instancias correctamente agrupadas.
- *Intrinsic methods*: Se aplican cuando no disponemos de la clase. Los métodos que conocemos intentan medir la cohesión **intra-cluster** y la separación **inter-cluster**

Inicialmente se pensó en utilizar un método de los que utilizan datos de los que se conoce la clase previamente. Pero siguiendo el criterio de que se trata de implementar un sistema capaz de explorar patrones comunes en conjuntos de datos de los que **NO** se conoce la clase, evaluar con un problema que no se ajusta a este hecho nos parece que no es una medida muy realista, dado que se conocen el número de clusters “óptimo” a priori y que atributos correlan mejor con la clase. Además en un contexto real en el que se intenta explorar un conjunto de datos para encontrar patrones similares, no se dispone de las predicciones o de la etiqueta de la clase, por lo que nos decidimos por analizar e implementar un método de evaluación intrínseco. El método escogido ha sido el coeficiente **Silhouette**.

Silhouette Coefficient

Es una combinación de las medidas de separación y cohesión:



Figura 1: Medidas de cohesión y separación

El coeficiente Silhouette s puede ser calculado para puntos independientes y para clusters. Para un punto individual, a = la distancia promedio de i a los puntos del mismo cluster; b = la distancia promedio de i a los puntos de los otros clusters.

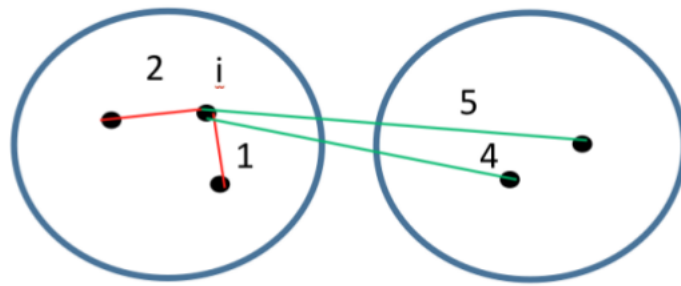


Figura 2: Coeficiente silhouette para un punto

Dependencias

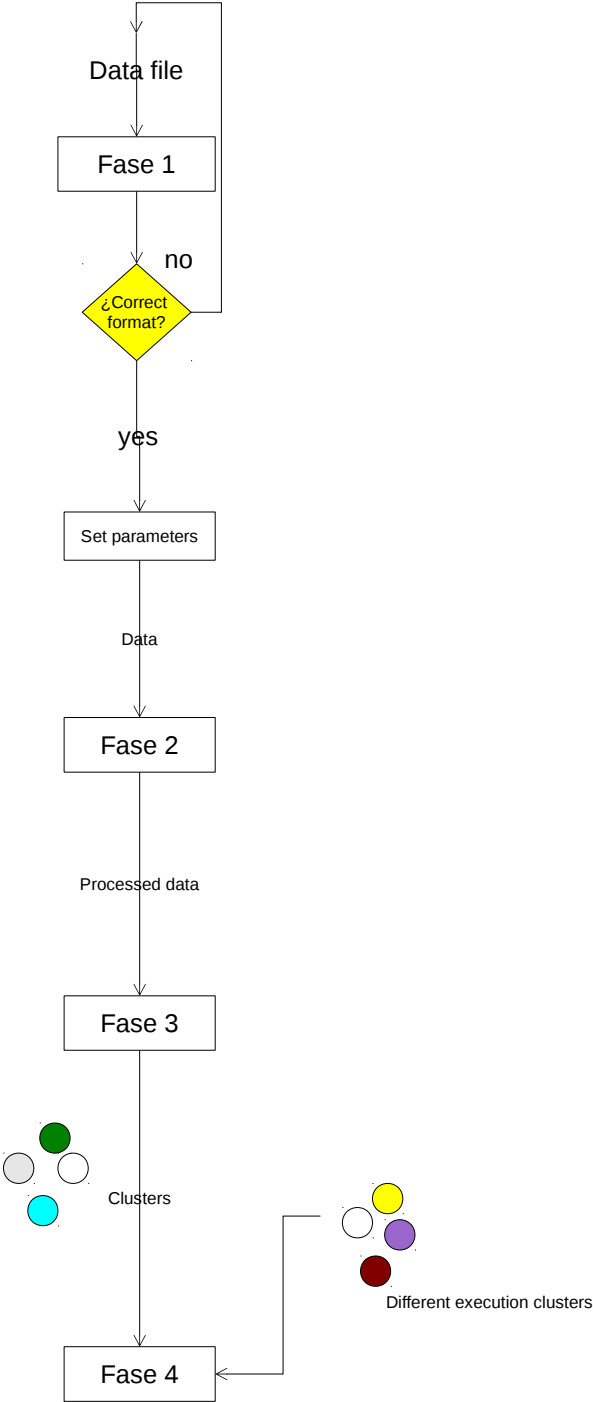


Figura 3: Dependencias del sistema

Para calcular el coeficiente del conjunto total de los clusters utilizamos la media del coeficiente para cada punto, la media de todos los puntos para calcular el del cluster, y la media de todos los coeficientes obtenidos para cada cluster.

4.1. Algoritmo en pseudocódigo

```
1  Let  $k$  be the number of clusters to partition the data set
2  Let  $X = x_1, x_2, \dots, x_n$  be the data set to be analyzed
3  Let  $M = m_1, m_2, \dots, m_k$  be the code-book associated to the clusters
4  Let  $dist(a, b)$  be the desired distance metric
5  Let  $B = B_{11}, B_{12}, \dots, B_{nk}$  be the temporary pertenence bit matrix
6
7  Ensure:  $C = C_1, C_2, \dots, C_k$  set of clusterized instances
8
9  Begin:
10
11  //randomly initialize the first centroids
12  for each  $m_j$ 
13     $m_j = randomsample(X)$ 
14  end
15
16  //assign dataset instances to each cluster generated by the centroids
17  for each  $x_n$ 
18     $B_{nj} = 1$  si  $\operatorname{argmin} dist(x_n, m_j) = m_j$  \foreach  $m_j$  si no  $B_{nj} = 0$ 
19  end
20
21  for each  $B_{nj}$ 
22    if  $B_{nj} == 1$ 
23       $C_j.add(x_i)$ 
24    end
25  end
26
27  //iterate the algorithm generatin new centroids based on previously clusterized instances until
  there are no changes between iterations
28  while changes in M do
29    for each  $m_j$ 
30       $m_{jnew} = calculatecentroid(C_j)$ 
31      if  $m_{jnew} == m_j$ 
32        changes = false
33      else
34        changes = true
35      end
36       $m_j = m_{jnew}$ 
37    end
38
39    for each  $x_n$ 
40       $B_{nj} = 1$  if  $\operatorname{argmindist}(x_n, m_j) = m_j$  \foreach  $m_j$  else  $B_{nj} = 0$ 
41    end
42
43    for each  $B_{nj}$ 
44      if  $B_{nj} == 1$ 
45         $C_j.add(x_i)$ 
46      end
47    end
48  end
49
50  return  $C = C_1, C_2, \dots, C_k$ 
51 end
```

5. Implementación

5.1. Análisis del conjunto de datos para decidir la conveniencia de la normalización

Debido a las dudas surgidas en torno a la normalización de los atributos y su conveniencia, consideramos adecuado para la tarea tratar de buscar algún método que fuese en alguna manera indicador de la utilidad de la normalización.

En un principio nuestro planteamiento se basaba en utilizar la media de la desviación típica de los valores de cada atributo con el objeto de poder analizar los rangos de las diferencias entre valores de cada atributo. Sin embargo la media por su cuenta no nos es útil, ya que por ejemplo, si la media es alta pero la desviación es baja, en realidad, puede no haber mucha variación en los rangos. Esto nos llevó a hallar el Coeficiente de Variación

$$C_V = \frac{\sigma}{\bar{x}} \quad (1)$$

De esta forma logramos un indicador más preciso sobre el rango que buscamos ya que nos indica la proporción de la variabilidad de las desviaciones, en lugar de la mera cantidad de desviación.

La motivación de este análisis viene dada más por el interés de hallar una forma de conocer la utilidad que pueda tener la normalización en un conjunto de datos, ya que objetivamente, el beneficio principal que puede tener es que si no afecta demasiado al resultado final, nos puede interesar más tener los atributos en su rango numérico inicial (p.ej.: es más visual ver un gráfico con edades entre 0 y 100 que entre 0 y 1).

Por otra parte, hemos tratado de hallar una cifra del Coeficiente de Variación que esté comunmente aceptada como baja, pero no hemos encontrado ninguna fuente fidedigna para ello. Por lo tanto, hemos decidido establecer una cifra pequeña (0,1) teniendo en cuenta que ante la posibilidad de que haya rangos muy variados, puede ser más conveniente normalizar.

Huelga decir que este método y sus resultados son experimentales, pese a tener cierta base empírica carecemos de la certeza sobre su eficacia, siendo nuestro objetivo principal investigar respecto a la normalización en lugar de simplemente aplicarla por estar aceptada como conveniente.

5.2. Problemas encontrados y soluciones adoptadas

5.2.1. Problema con la normalización de los atributos

En un comienzo, para normalizar, nuestra intención era aplicar la función Z-Score sobre los atributos de las diferentes instancias. Sin embargo, de esta forma se obtenían resultados fuera del rango de $[-1, 1]$, que es incorrecto para esta normalización. No se consiguió localizar el error, desconociendo si se trataba de una formulación incorrecta o un bug de programación. En lugar de ello se ha realizado una proyección lineal al intervalo $[0,1]$ y de esta forma se ha conseguido unificar los atributos en el mismo intervalo.

5.2.2. Problema de generación de excesivas instancias

Una vez estructurado todo el código y con todos los módulos programados, nos encontramos con el problema de que durante la ejecución, se generaban demasiadas instancias. Esto se debía a que en cada iteración del algoritmo, los clusters no eran reseteados correctamente, y cada vez que se añadía una nueva instancia en lugar de eliminar la que estaba presente en su lugar, desplazaba esta última y se insertaba en su posición. Para corregir esto sencillamente se modificó la función de añadir instancias para que esta sobrescribiese la presente, y además, se resetean los clusters en cada iteración.

5.2.3. Instancias repetidas en los gráficos

Al igual que en el anterior problema, nos encontramos con que a la hora de graficar, se marcaban más instancias de las que realmente había. La solución también pasó por realizar un reinicio de la matriz de bits en cada iteración del algoritmo.

5.2.4. Ausencia de representación para el cluster 0

Una vez incorporado el código necesario para la generación de gráficos su funcionamiento era correcto, pero no mostraba los datos del cluster 0. El problema radicaba en que a la hora de tratar la matriz de bits de pertenencia, el primer caso se trataba fuera de un bucle y por ello no se hacían las actualizaciones correctas. Una vez corregido esto el gráfico se crea correctamente.

6. Validación del *software*

6.1. Diseño del banco de pruebas

7. Análisis de resultados

7.1. Modificando inicializaciones

7.2. Modificando distancia Minkowski

7.3. Criterios de convergencia

7.3.1. Número fijo de iteraciones

7.3.2. Disimilitud entre *codebooks*

7.4. Distintas métricas

7.4.1. Manhattan

7.4.2. Euclídea

7.4.3. Minkowski

8. Clasificación supervisada respecto de

9. Conclusiones

9.1. Técnicas de clustering: motivación

Tal y como se ha descrito anteriormente en este documento, el propósito de la clasificación no supervisada es, opuestamente a la clasificación supervisada, **descubrir** información, en lugar de predecirla. Los campos y casos a los que es aplicable son múltiples y diversos: patrones de comportamiento en sociología, análisis genético, reconocimiento facial, etcétera. Todo área de conocimiento capaz de recopilar gran cantidad de datos es susceptible y puede beneficiarse de metodologías de clustering.

9.2. Conclusiones generales

Debido a la modularidad del software realizado, nos ha resultado relativamente sencillo encontrar los orígenes de los diversos errores y poder solucionarlos así como añadir varios módulos (p. ej.: generación de gráficos e informe) a lo largo del desarrollo, teniendo la estructura principal del código ya hecha. Además de esto, varias de las clases generadas son genéricas y pueden ser utilizadas en otros proyectos, como las funciones estadísticas de *Statistics.java* o la evaluación de clusters de *Evaluation.java*.

Por otra parte, la realización de esta práctica nos ha llevado a conocer en mayor profundidad el funcionamiento de un algoritmo de cierta complejidad, haciendo un uso mínimo de librerías externas. De no haber sido así, no se habría descubierto el algoritmo del coeficiente Silhouette, y aún menos haberlo implementado sin

acudir a código ajeno. También cabe destacar la capacidad para manejar diversas que se ha obtenido, así como el uso de varias funciones estadísticas para calcular si es apto normalizar o no (experimental).

9.3. Puntos débiles y propuestas de mejora

Dadas la complejidad y la diversidad de funciones implementadas en el software, ha habido ciertos aspectos en los que no ha sido posible centrarse tanto como cabía. Podría destacarse el rendimiento del mismo, que pese a ser correcto y no tomar demasiado tiempo con volúmenes grandes de datos, no ha sido puesto a revisión exhaustiva por lo que podría ser mejorado.

Además de esto, la cantidad de métodos de evaluación que se han podido implementar son limitados. Sin embargo no lo consideramos un punto débil destacable, ya que el algoritmo implementado, tras su análisis y diversas pruebas, nos ha parecido consistente y efectivo.

- Conclusiones a la vista de los resultados más relevantes.

10. Valoración subjetiva

1. ¿Has alcanzado los objetivos que se plantean?
2. ¿Te ha resultado de utilidad la tarea planteada?
3. ¿Qué dificultades has encontrado? Valora el grado de dificultad de la tarea.
4. ¿Cuánto tiempo has trabajado en esta tarea? Desglosado:

Coste temporal	
Diseño de software	1
Implementación de software	
Tiempo trabajando con Weka	
Búsqueda bibliográfica	
Informe	

5. Sugerencias para mejorar la tarea. Sugerencias para que se consiga despertar mayor interés y motivación en los alumnos.
6. Críticas(construktivas).

Referencias

[1] MOA