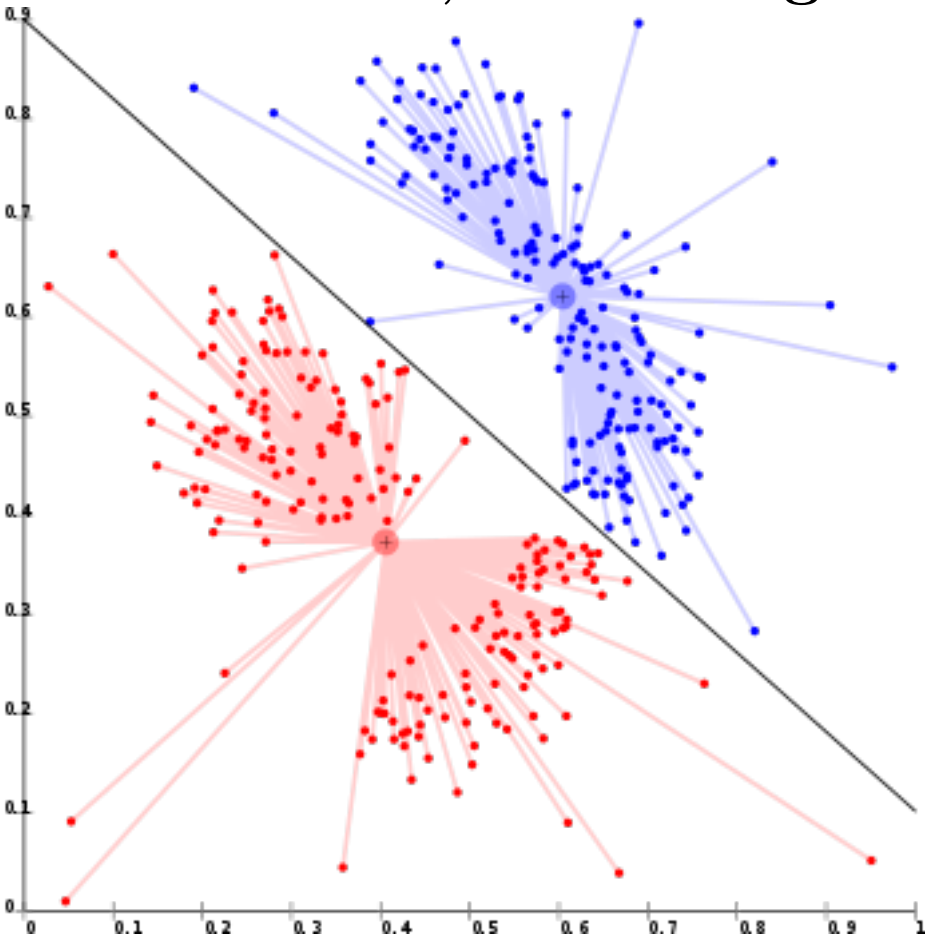
Práctica 6, Clustering



Minería de datos

Andoni Martín Reboredo
David Ramirez Ambrosi

19 de octubre de 2014

Índice general

Índice de figuras

Capítulo 1

Introducción

El presente documento constituye el resultado de la práctica realizada en base a la implementación del algoritmo de clasificación no supervisada **K-Means clustering**. Este algoritmo trata el agrupamiento de un conjunto de instancias en base a su proximidad con las demás instancias contenidas en el espacio de muestra proporcionado al algoritmo para su ejecución.

Dentro del algoritmo cabe el estudio de diferentes variaciones en los distintos parámetros de que dispone. Nosotros hemos considerado variaciones sobre dos parámetros, el método de cálculo de la distancia entre los distintos elementos que posee el cluster y la inicialización de los distintos clusters. Esta inicialización servirá como base de las sucesivas iteraciones que conforman el algoritmo.

1.1. Clasificación no-supervisada

La clasificación no supervisada es aquella que se lleva a cabo mediante el estudio de las diversas instancias que conforman el espacio de aplicación del algoritmo sin que estas instancias tengan que estar previamente clasificadas dentro de una clase [1].

Se trata de una técnica de exploración de los datos en la que se intentan detectar estas clases desconocidas. Dependiendo de el algoritmo de clasificación utilizado, el número de clases debe o no ser especificado. Por ejemplo, en el algoritmo en que se basa este trabajo debe ser especificado, sin embargo en técnicas de **clusterig jerárquico** no.

1.2. Objetivo

Esta práctica tiene como objetivo principal la comprensión de los procesos internos que realiza un algoritmo de agrupamiento cualquiera como puede ser el K-Means clustering. El aprendizaje se realizará de forma práctica a través de la implementación del algoritmo K-Means clustering junto con diversas opciones con las que realizar algunos pasos del mismo, como son el uso de métricas o inicializaciones del algoritmo diferentes. Estas variaciones requieren que el algoritmo sea entendido plenamente para poder hacer contribuciones que tengan utilidad para la realización del proceso.

Capítulo 2

Algoritmo

2.1. K-means, algoritmo principal

```
inicializar
divergencia = infinito

Mientras(numiteraciones <= iteracionesIndicadas AND divergencia < delta)
{
    centroides = centroidesNuevos

    calcularPertenencias
    centroidesNuevos = calcularNuevosCentroides

    calcularDivergencia(centroidesNuevos)
}
```

2.2. Subrutina inicialización

2.2.1. Inicialización aleatoria

```
Para cada dimensión
{
   mientras extraiga una instancia ya evaluada
   {
      extraigo una instancia nueva
   }
   añado la instancia extraida a las evaluadas
   establezco la instancia como centroide de un cluster
}
```

2.2.2. Pertenencia aleatoria

```
Mientras haya instancias que asignar
```

```
{
   Calculo un número de cluster aleatorio
   Extraigo la siguiente instancia
   Añado la instancia al cluster aleatorio
}
Calculo los centroides del cluster
```

2.2.3. División de espacio

```
Obtengo los rangos máximos y mínimos de cada subespacio
Mientras no haya creado k divisiones
{
   Mientras no haya establecido todos los atributos(recorrido los subespacios)
   {
      Divido el subespacio en K
      Asigno el centro del subespacio dividido correspondiente al índice del bucle
   }
   Añado el centroide resultante de dividir el espacio
}
```

2.2.4. Generación aleatoria de codewords

```
Obtengo los máximos y mínimos de cada dimensión
Para cada cluster
{
   Evaluo cada dimensión
   {
      Calculo un valor aleatorio para ese centroide en esa dimensión
   }
   Añado el centroide generado al cluster
}
```

2.3. Subrutina calcularPertenencias

```
Crear nuevos clusters

Para cada instancia
{
    Para cada centroide
    {
        distancia entre la instancia y cada cluster
        Guardamos los mínimos
    }
    Para cada centroide obtenido
        Guardamos la instancia en el cluster correspondiente al centroide
}
```

2.4. Subrutina calcularCentroides

```
Para cada cluster

   Calcular la instancia media

return nuevosCentroides
```

2.5. Subrutina calcularDivergencia

```
Para cada cluster

   Calcular la distancia entre el centroide antiguo y el nuevo

return divergenciaAcumulada
```

Capítulo 3

Diseño

Capítulo 4

Resultados experimentales

4.1. Banco de pruebas para la validación de software y resultados

4.2. Resultados

4.3. Clasificación supervisada respecto de otro software de refernecia

4.4. Análisis crítico y discusión de resultados

4.5. Rendimiento del software

Capítulo 5

Conclusiones

5.1. Motivación para la realización de clustering *Clustering*

5.2. Conclusiones de los resultados

5.3. Conclusiones generales

5.4. Propuestas de mejora

Bibliografía

[1] Alicia Pérez. Minería de datos, tema 9: Clasificación no-supervisada (clustering).

Capítulo 6

Valoración subjetiva

Alcance de objetivos

Utilidad de la tarea

Dificultad

Tiempo de trabajo

Sugerencias de mejora

Críticas