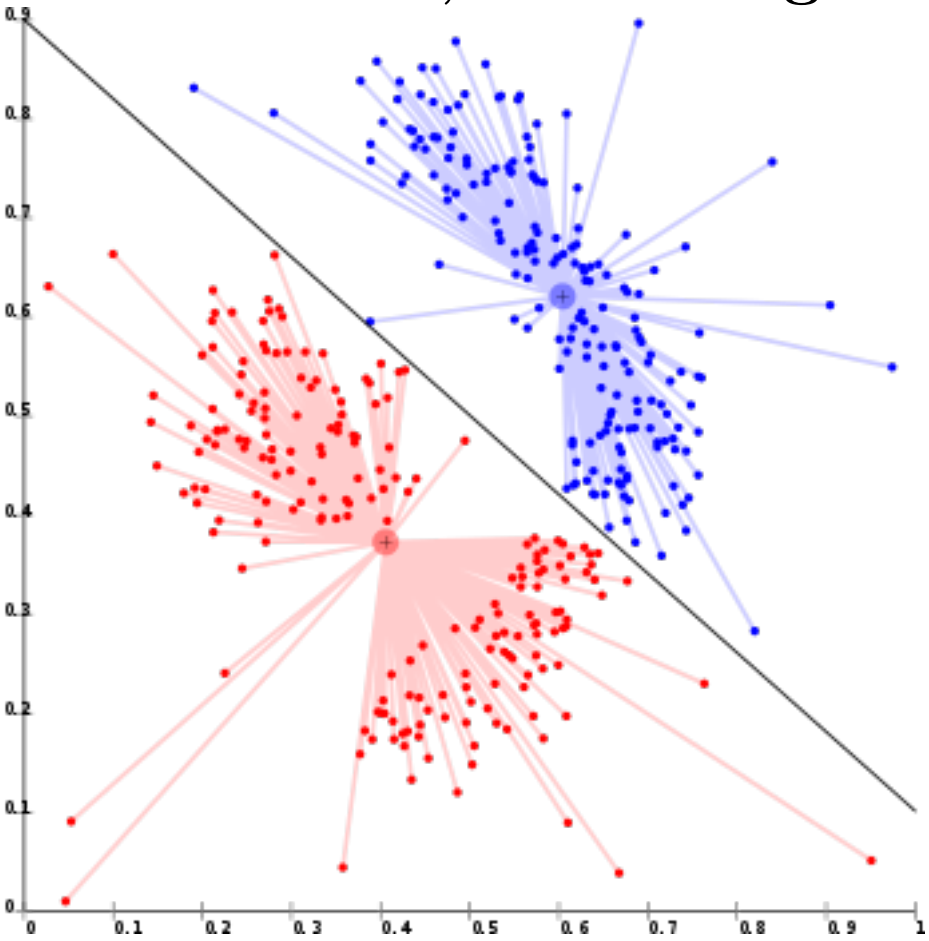
Práctica 6, Clustering



Minería de datos

Andoni Martín Reboredo
David Ramirez Ambrosi

15 de octubre de 2014

Índice general

1. Introducción **1**

1.1. Clasificación no-supervisada 1

1.2. Objetivo 1

2. Algoritmo **2**

2.1. K-means, algoritmo principal 2

2.2. Subrutina inicialización 2

2.2.1. Inicialización aleatoria 2

2.2.2. Pertenencia aleatoria 2

2.2.3. División de espacio 2

2.2.4. Generación aleatoria de codewords 2

2.3. Subrutina calcularPertenencias 2

2.4. Subrutina calcularCentroides 3

2.5. Subrutina calcularDivergencia 3

3. Diseño **4**

4. Resultados experimentales **5**

4.1. Banco de pruebas para la validación de software y resultados 5

4.2. Resultados 5

4.3. Clasificación supervisada respecto de otro software de refernecia 5

4.4. Análisis crítico y discusión de resultados 5

4.5. Rendimiento del software 5

5. Conclusiones **6**

5.1. Motivación para la realización de clustering *Clustering* 6

5.2. Conclusiones de los resultados 6

5.3. Conclusiones generales 6

5.4. Propuestas de mejora 6

6. Valoración subjetiva **7**

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 **KMeans 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.

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 comprehensión de los procedimientos

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

2.2.2. Pertenencia aleatoria

2.2.3. División de espacio

2.2.4. Generación aleatoria de codewords

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

Capítulo 6

Valoración subjetiva

Alcance de objetivos

Utilidad de la tarea

Dificultad

Tiempo de trabajo

Sugerencias de mejora

Críticas