Práctica de Laboratorio de Minería de Datos

Nombre: Carlos Patricio Pereira Paredes

Módulo: Décimo

Paralelo: "B"

1. **Objetivos**
   - Comprender la búsqueda de información dentro de la minería de datos
   - Realizar el análisis de minería de datos en la toma de decisiones con un ejemplo práctico.
   - Probar las herramientas utilizadas dentro de minería de datos

2. **Contenido Teórico(***información de la exposición y consultar el algoritmo de Kmedias***)**
   - **Definición de Minería de datos**
     Es un mecanismo de explotación, consistente en la búsqueda de información valiosa en grandes volúmenes de datos.
     La minería de datos se centrar en llenar la necesidad de descubrir el por qué, para luego predecir y pronosticar las posibles acciones con cierto factor de confianza para cada predicción
     Es el análisis de archivos y bitácoras de transacciones trabaja a nivel del conocimiento con el fin de descubrir patrones, relaciones, reglas, asociaciones o incluso excepciones útiles para la toma de decisiones.
   - **Algoritmo Kmedias**
     Es probablemente el algoritmo de agrupamiento más conocido. Es un método de agrupamiento heurístico con número de clases conocido *K*. El algoritmo está basado en la minimización de la distancia interna.

3. **Desarrollo**

   3.1. Enunciado del problema

   La empresa de software para Internet "Memolum Web" quiere extraer tipologías de empleados, con el objetivo de hacer una política de personal más fundamentada y seleccionar a qué grupos incentivar.
   Las variables que se recogen de las fichas de los 15 empleados de la empresa son:

   • **Sueldo:** sueldo anual en euros.
   • **Casado:** si está casado o no.
   • **Coche:** si viene en coche a trabajar (o al menos si lo aparca en el párking de la empresa).
   • **Hijos:** si tiene hijos.
   • **Alq/Prop:** si vive en una casa alquilada o propia.
   • **Sindic.:** si pertenece al sindicato revolucionario de Internet

• **Bajas/Año:** media del nº de bajas por año
• **Antigüedad:** antigüedad en la empresa
• **Sexo: H:** hombre, M: mujer.

Los datos de los 15 empleados se encuentran en el fichero "empleados.arff", que lo encuentran en la siguiente dirección web: www.dsic.upv.es/~cferri/weka. Se intenta extraer grupos de entre estos quince empleados.

3.2. ***Proceso de resolución del Problema***

Se utilizará el método de Cluster para ello acudimos a la ventana *Cluster*, luego seleccionaremos Choose para seleccionar el algoritmo SimpleKmeans, finalmente definimos 3 el número de cluster (**en este apartado agregaremos la ventana con los resultados que arroja**)

=== Run information ===

Scheme: weka.clusterers.SimpleKMeans -N 3 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation: empleados.txt-weka.filters.unsupervised.attribute.Remove-R1
Instances: 15
Attributes: 9
Sueldo
Casado
Coche
Hijos
Alq/Prop
Sindic.
Bajas/Año
Antigüedad
Sexo
Test mode: evaluate on training data

=== Model and evaluation on training set ===

kMeans
======

Number of iterations: 3
Within cluster sum of squared errors: 17.61279643369798
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (15)	Cluster# 0 (6)	1 (5)	2 (4)
Sueldo	21066.6667	29166.6667	16600	14500
Casado	No	No	Sí	Sí
Coche	Sí	No	Sí	Sí
Hijos	0.7333	0.1667	1.8	0.25
Alq/Prop	Alquiler	Alquiler	Prop	Alquiler
Sindic.	No	Sí	No	No
Bajas/Año	5.2667	6.1667	3.4	6.25
Antigüedad	8.2	8.3333	8.4	7.75
Sexo	H	M	H	H

Clustered Instances

0	6 (40%)
1	5 (33%)
2	4 (27%)

3.3. Análisis de los Resultados

Primera iteración

Con los resultados obtenidos lo que podemos observar es que en la primera iteración existen tres seis personas como muestra lo cual corresponde a un 40% de la población total, en el que se puede observar que las personas son solteras no poseen carro la tasa de personas con hijos es de 0.1667 hijos cada uno, no poseen domicilio propio pertenecen al sindicato de internet, en el mismo existe una baja de 6.1667 por año la mayoría tiene una antigüedad mayor a ocho años y son mujeres.

Segunda iteración

En la misma se puede observar que pertenece a cinco personas lo que corresponde a un 33% de la población total, en el que se puede observar que las personas son casadas, poseen carro la tasa de personas con hijos es de 1.8 hijos cada uno, poseen una propiedad, no pertenecen al sindicato de internet, en el mismo existe una baja de 3.4 por año la mayoría tiene una antigüedad mayor a ocho años y son Hombres.

Tercera iteración

Se puede observar que pertenece a cuatro personas lo que corresponde a un 27% de la población total, en el que se puede observar que las personas son casadas, poseen carro la tasa de personas con hijos es de 0.25 hijos cada uno, viven en un alquiler, no pertenecen al sindicato de internet, en el mismo existe una baja de 6.25 por año la mayoría tiene una antigüedad mayor a siete años y son Hombres.

4. Conclusiones

- Hemos comprendido el uso minería de datos aplicando técnicas de clasificación y de agrupación.
- Este tipo de tecnologías nos ayudado para afianzar el proceso de enseñanza-Aprendizaje.
- En resumen Minería de Datos se presenta como una tecnología innovadora, que presentan ventajas para la administración eficiente en la toma de deciciones.

5. Bibliografía

- Manual de prácticas de minería de datos usando el software WEKA, Abdelmalik Moujahid e Iñaki Inza, https://addi.ehu.es/bitstream/10810/4627/1/tr10-1.pdf

- Algoritmos De Aprendizaje: Knn & Kmeans, http://www.it.uc3m.es/jvillena/irc/practicas/08-09/06.pdf

- Algoritmos de Data Mining aplicados en la enseñanza basada en la web, http://www.palermo.edu/ingenieria/Pdf2010/CyT9/20.pdf

6. Licencia:

Litos

Firmado digitalmente por Litos
Nombre de reconocimiento (DN): cn=Litos, o=Cp, ou=Carlos Pereira, email=cpcubo@gmail.com, c=EC
Fecha: 2012.07.25 23:08:57 -05'00'