Clustering and data asociation

Alejandro Ye Xu

Universidad del Pais Vasco/ Eukal Herriko Unibertsitatea (December 12, 2019)

Objetivos

- Meta: Dados unos artículos de noticias (input: Texto sgml) encontrar relaciones entre dichas transacciones.
- Research questions:
  RQ1 ¿Porqué usar la técnica del clustering?
  1. ¿Jerárquico o k-means?
  2. ¿Cómo etiquetar el cluster?

  RQ2 ¿Qué son las reglas de asociación?
  1. ¿Diferencia entre las otras?
  2. Qué conclusiones se han llegado a obtener

Tarea y datos

Tarea

- Se propone extraer reglas de asociación de diferentes textos semi-etiquetados de noticias.
- Data source: [Reu,]

Caracterización de los datos: Se han utilizado las etiquetas dadas por los recopiladores de artículos, (Ej: Fecha, Tema, Sitios, Personas, Organizaciones, Intercambios, Empresas), más una etiqueta propia, caracterizada por el análisis del texto, en el título y cuerpo, y clusterizándo cada noticia. Data analysis graphically

Representación del texto

- **Pre-proceso** Se ha asegurando la codificación utf-8 y que el idioma sea Inglés. Asimismo, se ha tokenizado el texto mediante la librería NLTK (Natural Language Toolkit), quitando las palabras que no aportan información llamadas stop-words(Artículos, conjunciones...), lematizando y procesando sólo la raíz de dicha palabra.
- **Representation** Los datos para el clusterizado se han compuesto por TF-IDF
- **Lenguaje** Python junto con las librerías, nltk, numpy, sklearn y csv

Clustering Jerárquico

Este algoritmo agrupa datos dependiendo la distancia entre cada instancia e intenta que los datos que están dentro de un clúster sean lo más parecido. Se agrupan normamente de forma anidada en forma de arból, por lo que cuando son parecidos los nodos 'hijo', tienen el mismo 'padre'.

- Librería: from sklearn.cluster import AgglomerativeClustering

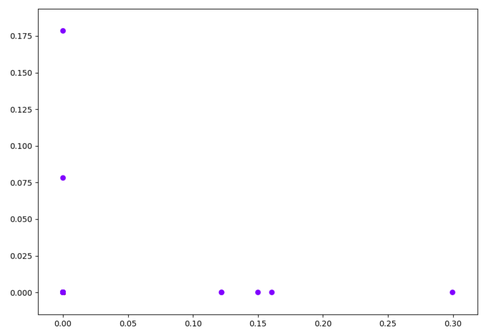

Figure 1:Posición de los clusters obtenidos con el algoritmo jerárquico

Clustering k-means

Este método de agrupamiento a diferencia del jerárquico parte de un conjunto de n instancias y las va particionando en k grupos.

- Librería: from sklearn.cluster import KMeans

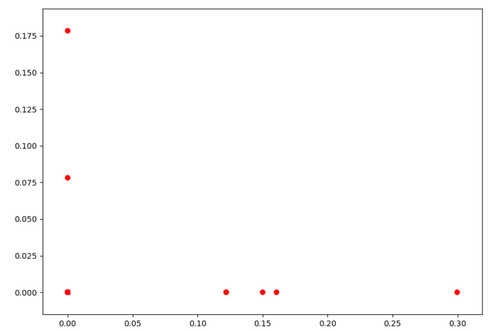

Figure 2:Posición de los clusters obtenidos con el algoritmo k-means

Lo curioso, es que el número de clústers óptimo para ambos es de cinco, con diferentes parámetros. (Con 5000 artículos) Sin embargo este dato no es tan curioso una vez explorado el texto, ya que parece haber más artículos de un tipo, relacionados con las finanzas o ganancias.

Uso del clustering en las reglas de asociación y difentes tipos

El clustering nos permite asignar una etiqueta más a los datos, pero los algoritmos nos devuelven un número. El cual hay que interpretar, por ello, en el código, se sugiere una manera sencilla de traducirlo y es observando las instancias en dicho clúster, obtener los temas de cada una de ellas y escoger las más repetidas. Tipos de reglas de asociación:

- Apriori: Usa búsqueda en anchura para calcular el soporte.
- Eclat: Usa una búsqueda en profundidad basada en la intersecciooón de un conjunto.
- FP-Growth: Encuentra patrones frecuentes, cuenta primero las ocurrencias entre los pares atributo-valor y después genera un árbol donde va añadiendo las transacciones

Reglas de asociación

Conceptos útiles:

- Support: El soporte es la frecuencia con la que aparece un conjunto en el dataset.

$$P(X,Y) \tag{1}$$

- Confidence: Indica la ocurrencia de que la regla haya acertado.

$$P(X|Y) = \frac{P(X,Y)}{P(X)} = \frac{P(X,Y)}{P(Y)} \tag{2}$$

- Lift: Es la proporcionalidad del support comparado con la indepencencia entre las variables.

$$P(X|Y) = \frac{P(X,Y)}{P(X) * P(Y)} \tag{3}$$

Discussion

- Los datos experimentales indican que los artículos analizados tienen poca relación, ya que, con un support=0,005 y confianza=0.05 solo se obtienen 3 reglas.
  cluster2_acq_earn -> usa (Esta regla indica que el clúster2 generado donde los articulos están relacionados con ganar o adquirir también está relacionado con Estados Unidos)4-mar-1987 -> cluster2_acq_earn (Esta regla asocia esa fecha con el cluster2)6-mar-1987 -> cluster1_earn (Esta regla asocia el cluster1 a esa fecha concreta)
- Se esperaba algo así, ya que, el texto en los artículos es escaso y muchas de las etiquetas proporcionadas están vacías. Si bajamos el support más obtendremos más reglas, con un support=0,001 se obtienen 89 reglas, y una de ellas es reagan -> usa, tiene sentido porque reagan fue el presidente de USA, gobernador de California desde 1967 hasta 1975 y presidente desde el '81 hasta el '89.

Conclusiones

- Extraer información de textos no es tan fácil como parece.
- Fortalezas: Interés, ganas y gusto a la programación
- Debilidades: Tiempo escaso al compaginar trabajo con estudios desde 1º.

Bibliography

[Reu,] Reuters-21578 text categorization collection data set.
https://archive.ics.uci.edu/ml/datasets/reuters-21578+text+categorization+collection.