Resumen de tkmins:

Lo que esta realizando es un argoritmo de clasterizacion en el que se basan en la modificación de kmeans para clasterizar transacciones.

Como lo hacen: en lugar de utilizar la distancia euclidiana estándar emplean la distancia de Jaccard que es mas adecuada para comparar conjuntos de elementos. Además redefinen el concepto de centroide para que represente mejor las características comunes de las transacciones dentro de un cluster.

Frase muy importante que define la clasterizacion de transacciones Transactional clustering is related to two classes of problems: clustering of categorical attributes and clustering of variable-length sets.

Archivo a leer S. Guha, R. Rastogi, and K. Shim. ROCK: A robust clustering algorithm for categorical attributes. In 15th International Conference on Data Engineering (ICDE 99), pages 512–521, Washington - Brussels - Tokyo, March 1999. IEEE. 175

En lugar de usar las distancias convencionales, el artículo utiliza la distancia de Jaccard, que mide la disimilitud entre conjuntos. Para representar los datos, los autores adoptan una representación compacta de vectores booleanos (que indican solo la presencia o ausencia de elementos, a diferencia de otros métodos como el de clustering de documentos que usan frecuencias de aparición).

Algo muy importante que esta haciendo este algoritmo es defininir las transacciones, lo primero que dice es que hay un conjunto de ítems y que en una transacción un item puede estar o no, luego hay muchas transacciones que son muchos conjuntos de itmes.

Supongamos que aia\_iai​ representa una URL web, y que I modela todas las páginas contenidas dentro de un servidor web. Un conjunto de ítems podría representar una sesión típica de un usuario en el servidor web, es decir, el conjunto de páginas que un usuario ha accedido dentro del servidor web dado.

Este enfoque es fundamental para abordar el clustering de datos transaccionales, donde cada transacción puede tener un número variable de elementos, como una compra que incluye diferentes productos.’

Definición 1

Lo primero que hace el paper es definir todos los conceptos de manera matemática relacionados con los cluster kmeans.

Luego te explica matemáticamente como cada objeto esta asociado a un único cluster

Luego te explica como maximiza ese algoritmo, para que todos los puntos pertenezcan al cluister que toca. Esto significa que se desea que, en cada cluster, los objetos estén lo más cerca posible del centroide del cluster.

Definición 2

Esta definición establece formalmente el concepto de centroide para un conjunto de objetos en un dominio U equipado con una función de distancia d: U x U-> R

La distancia mide la disimilitud entre dos elementos en el conjunto U.

En caso de datos transaccionales se suelen utilizar medidas basadas en superposición de elementos pe la distancia basada en la similitud Jaccard.

Para la definición de centroide explica lo siguiente; se busca encontrar un elemento v perteneciente al conjunto U que minimice la distancia de las sumasd e las distancias al cuadrado entre el u cada uno de los elementos que pertenecen a ese centroide

Hace una cosa que es elevar al cuadrado la distrncia, y explica que así se penalizan las las asignaciones en el que el objeto este muy alejado al candidato a centroide.

Entonces el centroide e el elemento que en promedio se encuentra mas cerca de todas las transacciones S

En espacios euclidianos y para datos numéricos, este centroide suele coincidir con el promedio. Sin embargo, en el caso de datos transaccionales (vectores de tamaño variable con valores categóricos), no es posible calcular un promedio aritmético convencional; en cambio, se debe emplear un método que, de forma análoga, identifique la transacción que mejor representa al conjunto.

Definición 3

Esta haciendo referencia a que cada transacción está formada por un conjunto de datos categóricos.

PI dice que es el conjunto de todas las posibres comunicaciones, por lo que cada transacion es un subconjunto de I que recordamos que i son todos lo elementos que puede haber en una transacción

2.1 Dissimilarity Measure; medida de desigualdad o disimilitud

Es muy importante esto, ya que esta definiendo como va a representar los datos transaccionales de manera no categórica, y va a ser una representación de vectores Booleanos, donde la longitud va a ser fija y cada posición del vector corresponde a un item.

I1​={a1​,a2​} → Vector: [1,1,0,0,0,0,0,0,0,0]

Por ejemplo esta transacion slo tiene el item 1 y el 2 por eso se representan a true esos dos valores

Con esta representación se podrían aplical medidas como la distancia de Minkowsky o el mismatch count donde miden la diferencia teniendo en cuanta tanto la presencia como ausencia de ítems

Pero estas medidas tiene limitaciones ya que no capturan completamente la intuición de la similitud entre transacciones. El problema clave es que al contar la ausencia de los ítems, transacciones muy separadas en la que la mayoría de los ítems no estén representadas usando estas medidas van a resultar parecidas ya que en cuanto a la ausencia se parecen

Este problema es similar a la clusterizacion de documentos, donde cada documento se representa como un vector de términos

Entonces en estos casos se usan medidas Ad-hoc, como Jaccard Coefficient(sj), donde mide la disimilitud contando el numero de términos en común.

Cosine Similarity evalua la orientación o angulo entre dos vectores capturando la direccionalidad de los términos donde estas medidas se pueden derivar una distancia simple. d(x,y)=1−s(x,y) donde esta trasformando similitud en disimilitud

Como para este problema vamos a usar la distancia de Jaccard la vamos a desglosar. Dada dos transacioones representadas por itemsets I y J se define :

∣I∩J∣ es la cantidad de ítems que ambas transacciones tienen en común.

∣I∪J∣ es la cantidad total de ítems presentes en al menos una de las dos transacciones.

Estos dos valores son dividios y restados a 1.

La interpretación cumple que a mayor intersección(mas ítems en comun) menos es la distancia es decir mayor es la similitud, y a mayor es la union mas ítems diferentes mayor es la distancia por lo que las transacciones son menos similares.

Cluster representative:

El representante del cluster es el elemento que minimiza la suma de las distancias cuadráticas respeto a dotas las transacciones que componen ese clúster. Por lo tanto este representante tiene que resumir las características de los componentes que forman este cluster. Por lo que debe capturar la esencia del grupo

Lo que defiuende aquí es que es muy costoso calcular estas distancias en transacciones

Usamos jaccard pero hay problemas relacuionados: no se puede buscar un representate ya que es inviable en tiempo polinomial si se exige que el representante sea un vector booleano además si se pudiera encontrar una solución esta solución no seria única

Para superar los problemas mencionados se propone usar una aproxiumacion, que se inspira en los representantes obtenidos de las distancias euclidianas y de conteo de desacuerdos. Se usa operaciones de unión e intersección

Lemma 1: el representante optimo de un cluster esta contenido en la unión de todos lo elementos de las transacciones del cluster. Esto traae un error, y esque podemos estar metiendo como representante un item que sale solo una vez por lo que que ese item este como rep, no es representativo del cluster. Además al hilo de esto, no se están teniendo en cuenta las frecuencias de aparición de los ítems,.

Que se propone para estos errores, usar la interseccion de las trancsaccones como una aproximación del representate del cluster. L as ventajas son que recoge únicamente aquellos ítems que aparecen en todas las transacciones del ckluster

Lemma 2: lo que defiende este lema es que la interseccion de todos los componentes de un subconjunto de transacciones es un representante optimo del cliuster.

Esto tiene partes negativas como que puede no estar teniendo en cuenta itemes que son importantes ya que aparecen en la mayoría de las transacciones no se están teniendo en cuenta, además en cluster muy poblados es probable que la interseccion sea vacia.

Cuidado porque este lema nos dice que la interseccion es siempre parte del representante pero no la define por completo

El lema 3 sostiene que el representante nunca puede ser vacio

Parte de la intersección: Se inicia con el conjunto de ítems que aparecen en todas las transacciones (lo cual garantiza una base común).

Refinamiento Iterativo: Se van añadiendo, de forma iterativa, los ítems más frecuentes (entre aquellos que aún no están incluidos) al representante, siempre y cuando ello contribuya a minimizar la suma de las distancias (o el error) dentro del clúster

El procedimiento greedy busca construir el representante del clustre de manera iterativa. Parte de un conjunto inicial como la interseccion y se van añadiendo de forma secuencial los ítems que aportan mayot frecuencia es decir aquellos que apoarecen con mas regularidad.

Aproximación adicional medienate uumbral. Esete umbral se introcue para evitar el elevasdo costo de ordenar los iutems en cada paso.

El parámetro umbral representa el grado de similitud eintra cluster que se desea alcanzar. Se interpreta como el porcentaje minim,o de ocurrencias que uinitem debe tener en las transacción del cluster para ser considera en la construcción del representante.

Definición 4

En estra definición formaliza la fgorma de calcular el representante de un conjunto de transacciones S mediante el uso de un umbral

Frecuencia de un item como el numero de transacciones en las que aparece el item lo que cuantifica cuan común es el item. Solo se incluyen en el representante aquellos ítems que apartecen en a menos un poircentaje Y de las transacciones del cluster.

Objetos fuera del cluster:

Bajo la definición de dJ, si dos transacciones tienen intersección vacía, su distancia es 1.

Por ello, un objeto que no comparte ningún ítem con ninguno de los representantes de los clústeres tendrá distancia 1 respecto a cada clúster y, en consecuencia, no se asigna a ninguno.

Estos objetos se consideran "trash" (basura) y, en el contexto del algoritmo, se tratan como posibles outliers o casos no representativos.

Para solucionar este problema se plantean tres soluciones: elegir de manera mas cuidadosa los centroides, por otro lado se puede ajustar iterativamente le numero de clústeres o el tamaño de la Y, o una de las mas interesantes es clusteriar de manera recursiva la basura, en la que se aplicaría nuevamente la técnica de clusterizacion sobre el conjunto de objetos de basura. Al seleccionar un conjunto aleatorio de representantes entre la basura y agrupar los elementos restantes, se pueden incorporar estos objetos en sub-clústeres, evitando que queden aislados debido a su alta disimilitud respecto a los clústeres principales.

Transactional kmeans

La idea principal es adaptar el tradicional algoritmo de kmeans al dominio de datos transaccionales iusando la distancioa de Jaccard para comparar transacciones donde tiene dos fases:

La primera es la creación de K+1 clústeres, donde se asignan las transacciones a los primeros k clusyeres según la distancia dj, es decir cada transacción se asocia al cluster cuyo representante es el mas cercano en términos de dj, se crea tmbn un cluistr adicional o basura que recoge aquellas transacciones que no han sido asignadas a ningún cluster

La segunda fase gestiona el cluster basura, la idea es aplicar recursivamente el mismo esquema de particviomn sobre el cluster basiuda dividiéndolo en cluster, de este modo se intentan agrupar aquellos outliers que son similares entre si. Cabe destacar que puede haber clusters con un solo elemento, que ocurre cuando algunos objetos que son muy disimiles a los demás se mantienen en el cluister de la basura hastga que mfinalmente se convierten en centro por si mismos.

Ahora va a realizar los experimientos, en esta fase me tengo que quedar en con quien lo compara, que datasets esta usando y que medidas usa para compararlos.

Lo primero compara el mismo algoritmo con dos técnicas para seleccionar el centroide, donde una usa repH y la otra repY

Se comopara con kmeans, y con ROCK lo que dice es quye kmeans tarda mucho en ejecutar y que ROCK es competitivo en datasets pequeños pero no en grances dada su complejidad cuadrática, indica que la versión repY es la mejor

Se observa que elk tiempo de ejecuciuon es inversamente proporcional al valor de Y

Lo siguiente es importante ya que evalua la alidad de la clusterizacion dicho de otra manera, que tan bien ha hecho el trabajo nuestro.

Medida de calidad basada en la Dissimilaridad

nfluencia de los Parámetros k y γ en la Población del Clúster Trash:

Los experimentos investigan cómo varían el número deseado de clústeres (k) y el umbral γ respecto al tamaño del clúster de basura. Se concluye que:

Un mayor valor de k (más clústeres principales) tiende a reducir el tamaño de la “basura”, ya que se capturan más patrones de similitud.

La elección adecuada de γ es crucial para equilibrar la homogeneidad interna de cada clúster y el costo computacional.

Medida Qd:

Se define como el promedio de los porcentajes promedio de las distancias intra-clúster respecto a la distancia total entre transacciones. Un problema de Qd es que, si hay una gran cantidad de objetos no clusterizados (el “trash”), éstos tienen, por definición, una distancia de 1 respecto a los representantes, lo que puede elevar artificialmente la similitud interna de los clústeres principales.

Medida F (F-measure) para datos sintéticos:

En los experimentos con datos sintéticos se asigna una etiqueta de clase a cada transacción (donde el número de clases C es un parámetro definido en el proceso de generación de datos). Para cada clúster, se calculan la precisión (p(i, j)) y el recall (r(i, j)) respecto a cada clase, y se obtiene el F-measure mediante

Comparación experimental:

Se comparan dos variantes del algoritmo propuesto (una basada en la heurística greedy y otra basada en el umbral γ) con el algoritmo ROCK y con versiones de K-Means que usan el vector medio como representante, empleando las medidas de distancia Jaccard y Coseno.

Se han utilizado tres conjuntos de datos: dos sintéticos (que varían en el número de ítems distintos y en la longitud promedio de las transacciones) y uno real basado en sesiones de usuarios web, las cuales tienden a ser muy heterogéneas.

A medida que aumentan la longitud de las transacciones y el número de ítems distintos, las transacciones resultan menos similares entre sí, lo que tiende a incrementar el tamaño del clúster trash.

Los resultados (resumidos en la Tabla 1 del estudio) muestran que, a pesar de ciertas sensibilidades (especialmente en los métodos basados en el umbral γ y ROCK, que dependen fuertemente de la elección del valor de γ), en general se obtienen resultados comparables en términos de calidad.

Caso practico:

Se usan los logs de un servidor proxy para analizar los patrones de la nacegacion de usuarios anónimos

Imagen que contiene Tabla

Descripción generada automáticamente