Métodos no supervisados y detección de anomalías: aspectos prácticos del método k-medias

UCO ONLINE

Aspectos prácticos del método k-medias

1. Introducción

El algoritmo k-medias es uno de los métodos de agrupamiento más populares por su sencillez, y ha servido de base para el desarrollo de muchos otros métodos. Una vez visto el funcionamiento del algoritmo, es interesante detenerse a analizar los aspectos que influyen en su comportamiento. El algoritmo k-medias es un método de partición, donde se intenta obtener una división no solapante de k grupos. La base fundamental de su funcionamiento es la asignación de instancias al centroide más cercano en función de una medida de distancia, habitualmente la Euclídea, con el objetivo de ir refinando la asignación de centroides hasta encontrar la mejor partición posible. Esta búsqueda de la mejor partición se traduce matemáticamente en la resolución de un problema de optimización: la minimización del criterio de la suma de errores al cuadrado (*Sum of squared error criterion*). Dicho criterio se formula como sigue:

$$SSE = \sum_{j=1}^{k}\sum_{i=1}^{N} \gamma_{ij}\|x_i - m_j\|^2$$

$$\gamma_{ij} = \begin{cases} 1 & si\ x_i \in c_j \\ 0 & si\ x_i \notin c_j \end{cases}$$

$$m_j = \frac{1}{N_j}\sum_{j=1}^{N} \gamma_{ij} x_i \qquad (1)$$

La primera fórmula representa la suma de las distancias entre las instancias y el centroide sobre la partición en k grupos ($C = \{c_1, \ldots, c_k\}$). La matriz de partición, formada por los valores γ_{ij}, establece la relación de asignación entre instancias (x_i) y grupos (c_j), de forma que solo se considere la distancia a las instancias pertenecientes al grupo al que representa el centroide (m_j). Como se vio en la lección anterior, las coordenadas del centroide se calculan como la media de todas las instancias asociadas al grupo (tercera fórmula).

La partición que minimiza este criterio se considera óptima y se le suele denominar la *partición de mínima varianza.* De la formulación de este criterio, basado en distancias, se infieren algunos de los supuestos que harán que k-medias resuelva satisfactoriamente o no el problema. En general, este criterio será adecuado en situaciones donde los grupos sean compactos y estén bastante separados. Por el contrario, este criterio es sensible a la presencia de valores anómalos (*outliers*), pues tenderán a dividir grupos grandes en muchos grupos pequeños. A continuación, se analizan estos y otros aspectos que influyen notablemente en el rendimiento del algoritmo k-medias.

2. Aspectos influyentes en k-medias

En este apartado se plantean una serie de aspectos que influyen en la ejecución y resultados que proporciona el algoritmo k-medias. Es importante conocer estos aspectos, así como posibles alternativas a la hora de aplicar este algoritmo en la práctica. De hecho, el análisis de algunos de estos problemas ha llevado al diseño de *variantes de k-medias*, esto es, nuevos algoritmos que modifican ligeramente algún paso del algoritmo original. Estos algoritmos se estudiarán en más detalle en otra lección de la siguiente semana, junto a otros métodos basados en partición.

2.1. Convergencia

El algoritmo k-medias es un método iterativo que sigue una estrategia "en escalada". Los llamados algoritmos "de escalada" (*Hill climbing*) toman en cada paso una decisión que solo permite mejorar la solución actual al problema. En el caso del criterio SSE, una nueva asignación de centroides debe reducir el error total en la partición. Este tipo de decisión puede llevar al algoritmo a estancarse en un óptimo local, cuando ninguna reasignación a partir del estado actual sea capaz de minimizar aún más el criterio SSE. Por tanto, un primer aspecto a considerar es que k-medias puede no converger a la solución óptima, debido principalmente a una mala selección de los centroides iniciales y a su propio proceso iterativo. Como solución a este problema se ha planteado el uso de otro tipo de optimizadores que eviten óptimos locales, como el enfriamiento simulado o los algoritmos genéticos. No obstante, estos procedimientos quedan fueran del alcance de este curso.

2.2. Inicialización

En su formulación básica, la selección de las instancias que se asignan inicialmente como centroides se realiza de forma aleatoria. Esto, sumado a la no garantía de convergencia, hace que el algoritmo k-medias pueda devolver diferentes particiones de un mismo conjunto de datos, aunque se establezca el mismo valor para k.

La sensibilidad a la inicialización puede abordarse con diversas estrategias. La primera y más sencilla consiste en ejecutar el algoritmo varias veces, de forma que cada una de ellas se inicializará con una asignación aleatoria diferente. Los resultados de cada ejecución nos pueden aportar conocimiento sobre la capacidad de convergencia del método. Si todas las ejecuciones terminan realizando una misma asignación de instancias a grupos a pesar de partir de una inicialización diferente, podemos estar más seguros de que el algoritmo converge.

Otros métodos más sofisticados analizan el conjunto de datos para determinar qué instancias son más interesantes como centroides iniciales. Entre ellas, el método propuesto por Kaufman y Rousseeuw consiste en determinar los centroides de forma sucesiva. El primero de ellos se corresponde con el punto ubicado lo más centrado posible respecto al resto de puntos. Este punto tendrá la menor distancia a todos los demás puntos. El siguiente punto que se selecciona como segundo centroide es aquel que reduce lo más posible el criterio de distancia a minimizar, y así sucesivamente. La idea general es seleccionar instancias representativas que tienen a su alrededor un número alto de instancias.

2.3. Selección del número de grupos

El algoritmo k-medias tiene un único parámetro, k, que indica el número de grupos a construir. Por tanto, k-medias asume que el usuario sabe cuántos grupos quiere encontrar, lo cual no suele ser cierto. Además, tampoco encaja con la filosofía del problema de agrupamiento, que nos decía que un método de agrupamiento nos ayudaría a encontrar los conceptos subyacentes a los objetos, aunque no supiéramos a priori cuáles son. Una estrategia simple para determinar el valor de k es ejecutar k-medias con varios valores de k, entre un mínimo y un máximo que estimemos representativo. Aquel valor de k que nos proporcione mejores resultados será el elegido. Como veremos al final de este documento, el análisis del valor de k es una de las primeras formas de validación de los resultados del algoritmo k-medias. Otros métodos de agrupamiento son capaces de autodeterminar el número óptimo de grupos como parte del proceso de agrupamiento.

2.4. Robustez

El algoritmo k-medias es sensible a la presencia de datos anómalos o ruido. En el proceso de asignación de instancias a grupos, k-medias considera por igual a todas las instancias del conjunto de datos. A todas y cada una de ellas le asignará un grupo, según su distancia al centroide, aunque esta distancia sea un valor excesivamente alto. Además, estas instancias "anómalas" van a participar en el cálculo de la nueva posición del centroide. Las instancias que contengan valores extremos en algunas de sus características van a distorsionar dicha posición, ya que la media de los valores deja de ser representativa.

Para evitar el primer efecto (asignar *outliers* a grupos), algunos métodos permiten dejar instancias sin asignar a ningún grupo. Para mitigar el segundo efecto (distorsión del centroide), algunos métodos optan por otras formas de calcular la posición como cambiar la media por la mediana, o tomar como centroide una instancia real del conjunto de datos. En cualquier caso, realizar un preprocesado de los datos también puede ser apropiado. En concreto, se debe analizar la idoneidad de normalizar las características y detectar (y en su caso eliminar) *outliers*.

2.5. Forma y densidad de los grupos

El algoritmo k-medias se basa en la distancia Euclídea para calcular la proximidad entre instancias. Esta distancia considera la diferencia de las magnitudes al cuadrado, lo que geométricamente implica que los grupos formados tenderán a tener una forma circular (o esférica en un espacio de más de dos dimensiones). El uso de otras medidas, como la Manhattan, puede ayudar a producir otro tipo de formas en los grupos. Aun así, otra característica de k-medias es su tendencia a crear grupos de tamaños similares, porque no considera la densidad de puntos dentro de cada grupo. Es decir, un punto se asigna al centroide más cercano, sin considerar cómo de cercano está al resto de puntos que ya hay en ese grupo. Sin una buena elección del valor k y en presencia de *outliers*, es posible que se generen grupos muy "dispersos" donde la distancia entre los puntos de un mismo grupo sea grande y, por tanto, los grupos sean poco representativos. Otros métodos basados en agrupación "suavizan" las restricciones sobre la morfología de los grupos y tienen en consideración la densidad de los puntos en el espacio.

2.6. Representatividad de los centroides

Los centroides en k-medias se calculan en cada iteración como el punto medio entre todas las instancias asignadas a un mismo grupo. Lo más probable es que el resultado de dicho cálculo no coincida con ninguna instancia real del conjunto de datos. En un espacio de características continuas, esto no tiene por qué ser un problema, ya que el valor medio calculado estará en el rango de valores posibles de cada característica. Sin embargo, en conjuntos de datos con características discretas, sobre todo sin son categorías y no números, no tiene por qué existir una definición o significado real del "punto medio". En general, es posible adaptar el algoritmo k-medias para que funcione sobre conjuntos de datos con características no numéricas. La alternativa más simple es sustituir la distancia Euclídea por otra función de distancia adecuada para el tipo de datos que se tenga. Respecto a la definición de los centroides, existen alternativas a k-medias que fuerzan a que el centroide sea una instancia real del grupo, aquella que mejor "represente" al grupo.

3. Selección del número de grupos

Aunque las medidas de validación de grupos se verán más adelante en el curso, la introducción del algoritmo k-medias ya nos hace plantearnos la necesidad de evaluar cómo se comporta este algoritmo ante diferentes valores de su parámetro k. A continuación, se describen tres técnicas que nos permiten comparar el resultado obtenido por k-medias (o cualquier otro algoritmo particional con k como parámetro) con el objetivo de decidir cuál es el mejor valor de k.

3.1. El método "del codo"

Al comienzo de la lección, se ha formulado la función de error (SSE) que utiliza internamente k-medias para tratar de encontrar la partición "óptima". El valor de esta función podemos calcularlo para cada partición resultante de ejecutar el algoritmo k-medias con un valor determinado de k. Puesto que queremos minimizar SSE, nos interesa un valor de k que nos ofrezca error pequeño. A la vez, es importante mantener un número de grupos reducido, ya que facilita la comprensión. Esta es la idea que subyace al método "del codo" (*Elbow method*). Este método considera SSE como una función dependiente de k. El objetivo es encontrar un compromiso entre el valor de k y el valor SSE asociado. Los pasos de este método se resumen a continuación:

1. Ejecutar el algoritmo de agrupamiento con diferentes valores de k, desde un mínimo hasta un máximo.
2. Para cada valor de k, calcular el valor SSE que proporciona la partición devuelta por el algoritmo.
3. Representar la curva de valores SSE en función del valor de k.
4. El punto de la curva donde haya un punto de inflexión se suele considerar como un indicador adecuado del número óptimo de grupos.

Se trata de un método visual muy sencillo de implementar, aunque no siempre sencillo de interpretar. Es posible que la curva no tenga una tendencia clara y no sea fácil identificar un valor adecuado de k. Además, este método se basa en SSE, una función de error que no todos los métodos de agrupamiento calculan de la misma manera.

3.2. El método "de la silueta"

Este método sigue una filosofía similar al anterior, pero es más general porque utiliza una medida de validación de grupos que es aplicable para el resultado de cualquier algoritmo. El coeficiente silueta (*Silhouette coefficent*) es una medida de calidad de un grupo que indica cómo de bien "encaja" cada elemento a su grupo asignado.

Aunque se estudiará en detalle en la segunda semana del curso, junto a otras medidas de validación, se presenta aquí brevemente. Cada instancia del conjunto de datos tiene asociado un valor silueta, que se calcula en base a dos términos: la disimilitud respecto a las instancias con las que comparte grupo (a_i en la ecuación 1), y la disimilitud con la instancia más cercana que no pertenece a su mismo grupo (b_i en la ecuación 1). Con estos dos valores, se calcula el coeficiente silueta para la instancia i como:

$$S_i = (b_i - a_i)/\max(a_i, b_i) \quad (1)$$

Si calculamos este valor para cada instancia y grupo, y tomamos la media entre los k grupos, el valor de k cuya partición alcance un valor medio más alto de este coeficiente será el mejor. El proceso sigue los mismos cuatro pasos que el método "del codo", pero sustituyendo el cómputo de SSE por el valor medio del coeficiente silueta en el segundo paso.

3.3. El método del estadístico "hueco"

A diferencia de los anteriores, este método tiene una base estadística, es decir, se basa en analizar la distribución de los datos y plantear una hipótesis sobre ella. Se trata de un método más reciente (publicado por R. Tibshirani, G. Walther y T. Hastie en 2001) y aplicable a cualquier método de agrupamiento.

Al igual que en los casos anteriores, se parte de la premisa de ejecutar el algoritmo con distintos valores del parámetro k. A continuación, el método compara una medida de variación intra-grupo, esto es, una medida en la que intervienen los puntos asignados a cada grupo, con los valores esperados si se presupone una distribución aleatoria de los datos. El número óptimo de grupos se asocia al valor del estadístico "gap" de mayor valor, pues significará que la partición es lo más diferente posible a una distribución aleatoria de los datos. El proceso consta de los siguientes cuatro pasos:

1. Ejecutar el algoritmo de agrupamiento sobre los datos originales, considerando varios valores de k entre un mínimo y un máximo. Calcular la variación intra-grupo para cada valor de k, W_k.
2. Generar B muestras de datos siguiendo una distribución aleatoria uniforme. Aplicar el algoritmo de agrupamiento sobre cada una de las muestras, variando el valor de k y calculando el valor $W_{k,b}$ como en el paso 1.

3. Calcular la estimación del estadístico como la desviación del valor W_k del valor esperado $W_{k,b}$ bajo la hipótesis nula (ecuación 1) y obtener su desviación estándar.

$$gap(k) = \frac{1}{B}\sum_{b=1}^{B} \log(W_{kb}{}^{*}) - \log(W_k) \qquad (1)$$

4. El número de grupos a elegir es el menor valor de k tal que la diferencia entre el estadístico en k y k+1 es inferior a la desviación estándar:

$$gap(k) \geq gap(k+1) - \sigma_{k+1}$$

Referencias

C. C. Aggarwal, C. K. Reddy (eds.). "Data Clustering: Algorithms and Applications". Chapman & Hall / CRC Press, 1ª edición, 652 páginas. 2014.

T. Hastie, R. Tibshirani, J. Friedman. "The Elements of Statistical Learning: Data Mining, Inference, and Prediction". Springer Series in Statistics, 2ª edición, 745 páginas. 2017.

G. James, D. Witten, R. Tibshirani, T. Hastie. "An Introduction to Statistical Learning with Applications in R". Springer Texts in Statistics, 1ª edición (7ª impresión), 426 páginas. 2017. Disponible en: https://www.statlearning.com/

A. Kassambara. "Practical Guide to Cluster Analysis in R". STHDA, 187 páginas. 2017.

S. Shalev-Shwartz, S. Ben-David. "Understanding Machine Learning: From Theory to Algorithms". Cambridge University Press, 1ª edición, 449 páginas. 2014.

R. Xu, D.C. Wunsch II. "Clustering". Wiley/IEEE Press, 1ª edición, 363 páginas. 2009.