Teoria-Resuelta-Examen-Enero-202...

user_2335920

Especialidad: Sistemas de Información

4º Grado en Ingeniería Informática

Escuela Técnica Superior de Ingenierías Informática y de Telecomunicación
Universidad de Granada

Accede al documento original

(1,5 ptos.) Justifica en que orden aplicar´ıas los siguientes métodos de procesado de datos: reducción del ruido, balanceo de clases por muestreo, selección de características e imputación de valores perdidos.

El orden ideal para aplicar estos métodos de procesado de datos sería el siguiente:

1. **Imputación de valores perdidos:** Este paso debe ser el primero, ya que los valores faltantes pueden generar problemas en los métodos posteriores. Por ejemplo, durante la selección de características o el balanceo de clases, los datos incompletos podrían influir negativamente en los resultados o incluso causar errores en los cálculos.
2. **Reducción del ruido:** Una vez que los datos están completos, es importante eliminar el ruido para mejorar la calidad de los datos y evitar que este interfiera en los procesos posteriores, como la selección de características. Si no se reduce el ruido antes, podríamos seleccionar características irrelevantes o afectar el balanceo de clases.
3. **Balanceo de clases por muestreo:** Este paso debe realizarse después de tener los datos completos y con menos ruido, ya que el balanceo depende de los datos disponibles. Si se realiza antes de imputar valores perdidos o reducir el ruido, se podrían generar instancias mal representadas que afecten al modelo.
4. **Selección de características:** Este paso debe ser el último porque depende de la calidad final de los datos procesados. Una vez que se han imputado valores, reducido el ruido y balanceado las clases, se puede elegir las características más relevantes para el problema, optimizando así el rendimiento del modelo.

Este orden asegura que cada etapa del procesado se realice sobre datos más fiables, completos y representativos, maximizando la eficacia de los pasos posteriores.

(2 ptos.) Compara justificadamente ventajas y desventajas de los siguientes algoritmos de clustering: K-Means, DBSCAN, Mean Shift, Jerárquico Aglomerativo Ward y BIRCH

1. K-Means

Ventajas:

- Es eficiente en términos computacionales, especialmente en grandes conjuntos de datos con características numéricas.
- Fácil de implementar y comprender.
- Se puede ajustar fácilmente el número de clusters mediante el parámetro kkk.
- Funciona bien cuando los clusters son esféricos y de tamaño similar.

Desventajas:

- Requiere especificar el número de clusters (kkk) de antemano.
- Sensible a los valores atípicos, que pueden desviar el centroide.

- No es efectivo con clusters de forma no esférica o densidades variables.
- Puede converger a mínimos locales, dependiendo de la inicialización.

2. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Ventajas:

- No requiere especificar el número de clusters de antemano.
- Puede encontrar clusters de forma arbitraria (no solo esféricos).
- Identifica y etiqueta automáticamente los valores atípicos como ruido.
- Es útil para datos con densidad variable.

Desventajas:

- Depende mucho de los parámetros ϵ\epsilonϵ (radio) y minPts\text{minPts}minPts (mínimo número de puntos en un cluster), que pueden ser difíciles de ajustar.
- No funciona bien si los clusters tienen densidades muy distintas.
- Puede ser costoso en términos computacionales para conjuntos de datos grandes.

---

3. Mean Shift

Ventajas:

- No requiere especificar el número de clusters, ya que determina automáticamente el número óptimo basado en los modos de la densidad.
- Puede encontrar clusters de forma arbitraria.
- Es robusto frente a valores atípicos.

Desventajas:

- Es computacionalmente costoso, especialmente para grandes conjuntos de datos, debido al cálculo iterativo de densidades.
- El tamaño del kernel (bandwidth\text{bandwidth}bandwidth) es un parámetro clave y puede ser difícil de determinar.
- No escala bien con datos de alta dimensionalidad.

4. Jerárquico Aglomerativo (Ward)

Ventajas:

- No requiere especificar el número de clusters de antemano, ya que produce un dendrograma que permite visualizar la estructura jerárquica.
- Es efectivo para conjuntos de datos pequeños y medianos.
- La metodología de Ward minimiza la varianza dentro de los clusters, produciendo resultados más compactos y homogéneos.

Desventajas:

- Es computacionalmente costoso (O(n2)O(n^2)O(n2)) para grandes conjuntos de datos.
- Las decisiones de fusión en etapas tempranas no se pueden revertir, lo que puede llevar a una jerarquía incorrecta si hay ruido o valores atípicos.
- No es adecuado para datos no esféricos o clusters de tamaños desiguales.

---

5. BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies)

Ventajas:

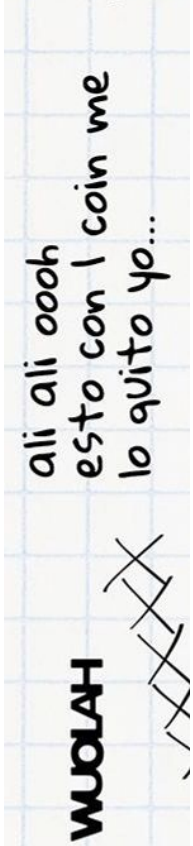
- Es eficiente en términos de memoria y tiempo, incluso para grandes conjuntos de datos, gracias a la representación compacta mediante árboles CF (Clustering Feature).
- Permite realizar clustering incremental, adaptándose a datos dinámicos.
- Escalable y capaz de manejar ruido si se configura adecuadamente.

Desventajas:

- El rendimiento depende del tamaño del umbral y del número de clusters en el árbol CF, lo que puede requerir ajuste manual.
- Menos efectivo en datos con clusters no esféricos o densidades muy variables.
- Puede perder detalles finos al resumir datos en los nodos del árbol CF.

Comparación general y usos recomendados:

- **K-Means**: Ideal para datos simples y cuando se sabe el número de clusters.
- **DBSCAN**: Excelente para clusters de forma irregular y presencia de ruido.
- **Mean Shift**: Útil para clusters arbitrarios y cuando no se quiere predefinir el número de clusters, pero costoso.
- **Jerárquico Ward**: Bueno para conjuntos pequeños y visualización jerárquica de relaciones entre clusters.
- **BIRCH**: Óptimo para conjuntos grandes con limitaciones de memoria, aunque menos preciso para estructuras complejas.



(1,5 ptos.) Explica qué inconvenientes puede presentar la confianza para valorar la fiabilidad de una regla de asociación y propón una alternativa que palíe esos inconvenientes.

Inconvenientes de usar la confianza para valorar la fiabilidad de una regla de asociación:

La **confianza** mide la probabilidad condicional de que un ítem BBB esté presente en una transacción dado que AAA ya está presente, es decir:

Confianza=P(B∣A)=Soporte(A∩B) / Soporte(A)

Aunque es una métrica popular, presenta los siguientes inconvenientes:

1. **Dependencia del soporte de BBB:**
   La confianza no tiene en cuenta la frecuencia individual de BBB. Si BBB es muy frecuente en el conjunto de datos (un ítem "popular"), la confianza puede ser alta incluso si la relación entre AAA y BBB es débil o espuria.
2. **Incapacidad de detectar reglas sin valor informativo:**
   Una regla como A→BA \to BA→B con alta confianza puede ser engañosa si BBB ocurre casi siempre, independientemente de la presencia de AAA. En este caso, la confianza no refleja la verdadera asociación causal entre los ítems.
3. **No diferencia entre correlación positiva y negativa:**
   La confianza solo mide la probabilidad condicional, pero no indica si AAA y BBB están correlacionados positivamente, negativamente o si simplemente coexisten por azar.

Alternativa: El Lift como métrica complementaria

El **lift** mide la fuerza de la relación entre AAA y BBB, comparando su coocurrencia observada con la esperada si AAA y BBB fueran independientes:

Lift(A→B)=Confianza(A→B) / Soporte(B)

Ventajas del Lift:

1. **Identifica independencia:**
   Un Lift=1\text{Lift} = 1Lift=1 indica que AAA y BBB son independientes, mientras que valores mayores o menores que 1 reflejan una relación positiva o negativa, respectivamente. Esto evita falsas interpretaciones de confianza alta debidas a ítems populares.
2. **Detecta asociaciones más significativas:**
   Al normalizar por el soporte de BBB, el lift evalúa si BBB ocurre más frecuentemente de lo esperado en combinación con AAA.
3. **Es más robusto frente a ítems con alta frecuencia:**
   Reduce el sesgo hacia ítems frecuentes al considerar la probabilidad base de BBB.

Conclusión:

Aunque la **confianza** es útil como métrica inicial, no es suficiente para valorar la fiabilidad de una regla de asociación. Complementarla con el **lift** permite identificar asociaciones significativas y evitar falsas interpretaciones causadas por ítems populares o asociaciones débiles.