Prueba Técnica: Minería de contrastes

La prueba consiste en implementar un flujo de minería de datos sobre un conjunto de datos con información salarial de una organización. El conjunto de datos está disponible para descargarse en este enlace.

La minería de contrastes (*Contrast Set Mining*, CSM) es una técnica de minería de datos con el propósito de encontrar conjuntos de características que señalen diferencias significativas entre grupos.

Antecedentes

A continuación se describe de forma general los conceptos principales detrás de esta técnica, una versión modificada de la propuesta realizada en el siguiente artículo de investigación: https://dl.acm.org/doi/pdf/10.1145/312129.312263

Un conjunto de contraste está dado por un conjunto de pares atributo-valor en conjunción; por ejemplo: `(gender=Female) ∧ (jobTitle=Software Engineer)`. Los grupos están dados por una variable de grupo, donde cada valor distinto que toma dicha variable conforma un grupo diferente.

En primera instancia se define una variable de grupo como entrada del algoritmo. Se considera que existe un grupo por cada valor diferente que toma la variable. Por ejemplo, cuando se selecciona la variable `seniority` como variable de grupo, cada uno de los valores del rango entre 1 y 5 conforman un grupo

Para evaluar un conjunto de contraste, se determina el soporte de cada grupo, el cual es el porcentaje de registros en el grupo que cumplen con el conjunto de contraste. En términos probabilísticos, es la probabilidad condicional de pertenecer al conjunto de contraste dado un grupo, formalmente, `P(cset | grupo)`.

Una vez se conoce el soporte de cada grupo, se determina la desviación, que es la diferencia entre el máximo y mínimo de soporte observado entre los grupos. Antes de ejecutar el algoritmo es necesario definir un umbral de desviación mínima (MINDEV), el

cual señala el valor mínimo que debe tener la desviación de un conjunto de contraste para ser considerado como interesante.

Para que un conjunto de contraste sea **viable** es necesario cumplir con los siguientes criterios:

- **Grande.** El conjunto de contraste debe superar la desviación mínima.
- **Significativo.** La asociación entre la variable de grupo y el conjunto de contraste debe ser estadísticamente significativa. Para determinar esto es necesario aplicar la prueba estadística chi2 a una tabla de contingencia donde las filas son la pertenencia al conjunto de contraste y las columnas los grupos. Use 0.05 como umbral para comparar el p-valor.
- **Productivo.** Un conjunto de contraste es productivo si todos sus subconjuntos propios son también conjuntos de contraste viables.

Tareas a realizar

Análisis exploratorio de datos

Plantee una descripción exploratoria del conjunto de datos, incluyendo al menos dos visualizaciones de datos.

Pre-procesamiento de datos

La minería de contrastes trata con datos exclusivamente categóricos, por lo que es necesario aplicar alguna técnica de transformación de datos para convertir variables numéricas en variables categóricas.

Minería de contrastes

Implemente un algoritmo de minería de contrastes que recupere todos los conjuntos de contraste viables en los datos proporcionados. Ejecute el algoritmo una vez por cada variable del conjunto de datos, tomándola como variable de grupo.

Resultados

Los resultados de la prueba están compuestos por el repositorio en Github así como de un reporte en formato PDF con los resultados obtenidos. A continuación se señalan los elementos a considerar en cada uno de ellos:

Repositorio

Repositorio en GitHub que incluya:

- Código utilizado.
- Archivo README que describa los pasos a seguir para ejecutar el flujo de minería.
- Archivo requirements.txt con dependencias requeridas por el código.
- Tabla de resultados que contenga exclusivamente los conjuntos de contraste viables. Los resultados de las ejecuciones por cada variable de grupo deben estar consolidados en esta tabla.

Reporte

Escriba un reporte en el cual se incluyan los siguientes elementos. Puede incluir información adicional si la considera necesaria. Considere los siguientes puntos:

1. Análisis exploratorio de datos.
2. Descripción de la solución.
3. Resultados obtenidos.
4. Interpretación de resultados.
5. Conclusiones.

Descripción de la solución

- Explique el método aplicado para la transformación de variables.
- Diagrama a bloques que describa su implementación del algoritmo.
- Justificación de su elección de valor de MINDEV.

Resultados obtenidos

Construya una tabla que consolide los resultados de todas las ejecuciones. Por cada ejecución, señale:

- Variable de grupo
- Total de conjuntos de contraste explorados
- Total de conjuntos de contraste viables encontrados
- Conjunto de contraste con la mayor desviación y su valor, así como los grupos de mayor soporte y menor soporte y sus valores.

Interpretación de resultados

Seleccione dos variables de grupo y desarrolle, para los resultados de cada ejecución, una interpretación de los conjuntos de contraste que considere más relevantes entre los resultados obtenidos.

Referencias

Bay, Stephen D., y Michael J. Pazzani. «Detecting Change in Categorical Data: Mining Contrast Sets». En Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 302-6. San Diego California USA: ACM, 1999. https://doi.org/10.1145/312129.312263.