ÁREA 4. SOLUCIONES DE CÓMPUTO INTELIGENTE

SUBÁREA 4.2 MINERÍA DE DATOS

TEMAS IMPORTANTE:

**Descubrimiento de Conocimiento en Bases de Datos:**

La minería de datos es una disciplina que se enfoca en descubrir patrones, tendencias y relaciones en grandes conjuntos de datos. Se utiliza para extraer información valiosa y conocimiento útil a partir de datos que de otra manera podrían ser difíciles de entender o analizar. La minería de datos utiliza técnicas de inteligencia artificial, aprendizaje automático y estadísticas para explorar y analizar los datos.

Knowledge Discovery in Databases (KDD) es un proceso que se utiliza para descubrir conocimiento útil a partir de grandes conjuntos de datos. Es un enfoque sistemático que combina técnicas de minería de datos, aprendizaje automático, estadística y visualización de datos para explorar y analizar los datos de manera eficiente.

El proceso de KDD consta de varias etapas:

* Selección de datos: se seleccionan los datos relevantes para la tarea de descubrimiento de conocimiento.
* Preprocesamiento de datos: se limpian y transforman los datos para asegurar que sean consistentes y estén en un formato adecuado para su análisis.
* Transformación de datos: se aplican técnicas de transformación de datos para reducir la dimensión de los datos o para convertirlos a un formato más adecuado para el análisis.
* Minería de datos: se aplican técnicas de minería de datos, aprendizaje automático y estadísticas para descubrir patrones, tendencias y relaciones en los datos.
* Evaluación de patrones: se evalúan los patrones descubiertos para determinar su relevancia y utilidad.
* Presentación de resultados: se presentan los resultados de manera clara y accesible para su uso en la toma de decisiones.

El objetivo de KDD es descubrir conocimiento útil y no trivial a partir de grandes conjuntos de datos, lo que puede ayudar a mejorar la toma de decisiones y a generar valor para las organizaciones. La aplicación de técnicas de KDD ha sido muy exitosa en una amplia variedad de campos, incluyendo el marketing, la detección de fraudes, la medicina, la ciencia y la ingeniería.

Diagrama

Descripción generada automáticamente

**Datos:**

Existen diferentes tipos de datos que se utilizan en la informática y en el análisis de datos, entre ellos:

* Datos numéricos: son datos que representan valores numéricos. Estos datos se pueden dividir en dos subtipos: datos discretos (por ejemplo, el número de hijos en una familia) y datos continuos (por ejemplo, la altura de una persona).
* Datos categóricos: son datos que representan categorías o etiquetas. Estos datos se pueden dividir en dos subtipos: datos nominales (por ejemplo, el color de los ojos) y datos ordinales (por ejemplo, la clasificación de una carrera).
* Datos de texto: son datos que representan texto, ya sea en forma de documentos, correos electrónicos, tweets, etc.
* Datos de imagen: son datos que representan imágenes, ya sea en forma de fotografías, dibujos, gráficos, etc.
* Datos de audio: son datos que representan sonido o música, ya sea en forma de grabaciones, canciones, efectos de sonido, etc.
* Datos espaciales: son datos que representan información espacial, ya sea en forma de mapas, coordenadas geográficas, modelos tridimensionales, etc.
* Datos de tiempo: son datos que representan información temporal, ya sea en forma de series de tiempo, eventos cronológicos, etc.

Los conceptos de dataset (conjunto de datos), datos crudos (raw data), database (bases de datos), etc.

* Dataset (conjunto de datos): es un conjunto de datos estructurados que se utilizan para entrenar modelos de aprendizaje automático, para realizar análisis estadísticos o para cualquier otro fin. Un dataset puede incluir información de diferentes tipos, como datos numéricos, categóricos, de texto, de imagen, etc.
* Datos crudos (raw data): son datos sin procesar y sin estructura que se obtienen directamente de una fuente, como sensores, sistemas de monitoreo, encuestas, etc. Estos datos pueden requerir una limpieza y transformación antes de poder ser utilizados para análisis o entrenamiento de modelos.
* Base de datos (database): es un conjunto organizado de datos que se almacenan y se acceden a través de un sistema de gestión de bases de datos (DBMS). Las bases de datos pueden incluir uno o varios datasets y se utilizan para almacenar información de forma estructurada, para realizar consultas y para generar informes.
* Data warehouse (almacén de datos): es una base de datos centralizada que se utiliza para almacenar grandes cantidades de datos de diferentes fuentes y sistemas. Un data warehouse se utiliza para realizar análisis y consultas complejas para la toma de decisiones empresariales.
* Data lake (lago de datos): es una infraestructura de almacenamiento de datos que permite almacenar grandes cantidades de datos de diferentes fuentes en su formato original, sin necesidad de transformarlos. Un data lake se utiliza para realizar análisis avanzados y para el entrenamiento de modelos de aprendizaje automático.
* Big Data:Se refiere a conjuntos de datos extremadamente grandes y complejos que no pueden ser procesados y analizados mediante herramientas tradicionales de procesamiento de datos. Estos conjuntos de datos son generados por diversas fuentes, como sensores, redes sociales, dispositivos móviles, transacciones en línea, entre otros.

**Preprocesamiento y Transformación:**

El preprocesamiento de datos es una etapa crucial en la minería de datos y el análisis de datos, ya que los datos pueden contener errores, ruido, valores atípicos, datos faltantes y otras irregularidades que pueden afectar negativamente el análisis y la calidad de los resultados. Algunas operaciones comunes de preprocesamiento de datos son:

* Limpieza de datos: se refiere a la eliminación de errores y datos duplicados en los conjuntos de datos.
* Remover ruido: se refiere a la eliminación de datos que son inconsistentes, irrelevantes o que no se ajustan a las tendencias y patrones generales del conjunto de datos.
* Remover outliers: se refiere a la eliminación de datos que se encuentran muy lejos de la mayoría de los datos y que pueden sesgar los resultados del análisis.
* Imputación de datos faltantes: se refiere a la estimación o imputación de valores para los datos faltantes en un conjunto de datos.
* Transformación de datos: se refiere a la transformación de los datos originales en una forma más adecuada para el análisis, como la normalización o la estandarización.

Estas operaciones de preprocesamiento pueden realizarse mediante herramientas y técnicas específicas, como algoritmos de limpieza de datos, técnicas de imputación de datos, técnicas de detección de outliers, entre otras.

**Trasformación:**  
El proceso de transformación de KDD, también conocido como proceso de preprocesamiento de datos, es una de las etapas principales del proceso de descubrimiento de conocimiento en bases de datos (KDD).

El proceso de transformación de KDD se divide en varias etapas, entre las que se encuentran:

* Selección de datos: En esta etapa, se seleccionan los datos relevantes para el análisis a partir de una base de datos o de varias fuentes de datos.
* Preprocesamiento de datos: En esta etapa, se lleva a cabo la limpieza de datos, que implica la eliminación de datos ruidosos, datos redundantes y datos faltantes. También se realiza la integración de datos, que implica la combinación de datos de diferentes fuentes en un solo conjunto de datos coherente.
* Transformación de datos: En esta etapa, se transforman los datos en un formato que sea adecuado para el análisis. Esto puede incluir la normalización de datos, la estandarización de datos, la discretización de datos y la codificación de variables categóricas.
* Selección de características: En esta etapa, se seleccionan las características más relevantes para el análisis y se eliminan las características irrelevantes.
* Reducción de datos: En esta etapa, se reduce la dimensionalidad de los datos mediante técnicas como el análisis de componentes principales (PCA) o el análisis de discriminante lineal (LDA).

**Selección:**El proceso de selección de KDD se refiere a la etapa en la que se seleccionan las características más relevantes del conjunto de datos para el análisis posterior. Esta etapa tiene como objetivo reducir la dimensionalidad de los datos, es decir, eliminar características irrelevantes y redundantes que no aportan información significativa al análisis y, por lo tanto, simplificar el modelo.

El proceso de selección de características en KDD se puede dividir en tres etapas principales:

* Extracción de características: Esta etapa implica la identificación de características relevantes que pueden ser utilizadas en el análisis posterior. Se pueden utilizar técnicas de minería de datos y estadísticas para extraer características significativas.
* Selección de características: En esta etapa se seleccionan las características más relevantes identificadas en la etapa anterior. Se pueden utilizar diferentes técnicas de selección de características, como la selección basada en filtros, la selección basada en wrappers y la selección basada en incrustaciones.
* Evaluación de características: En esta etapa se evalúan las características seleccionadas para determinar su efectividad en el análisis posterior. Se pueden utilizar técnicas de validación cruzada y otras técnicas de evaluación para evaluar las características seleccionadas.

El proceso de selección de características es una etapa importante en el proceso de KDD, ya que puede mejorar la precisión y eficiencia del modelo, reducir la complejidad del modelo y facilitar la interpretación de los resultados del análisis.

**Similitud y Distancias:Similitudes:**

La minería de datos no se limita a calcular la similitud entre diferentes datos, sino que se trata de un proceso más amplio que implica descubrir patrones y conocimientos útiles a partir de grandes conjuntos de datos.

Si bien es cierto que el cálculo de similitud entre diferentes datos es una tarea común en la minería de datos, como por ejemplo el cálculo de similitud entre vectores de características o entre conjuntos de elementos, la minería de datos también implica otras tareas, como la clasificación, la regresión, la asociación, la segmentación y la detección de anomalías.

En la minería de datos, se utilizan técnicas y algoritmos de aprendizaje automático, estadísticas y visualización de datos para descubrir patrones y conocimientos útiles a partir de grandes conjuntos de datos. Estos patrones y conocimientos pueden ser utilizados para tomar decisiones informadas y mejorar la toma de decisiones en diferentes áreas, como la medicina, la ingeniería, el marketing y la seguridad, entre otras.

Efectivamente, dentro de las medidas de similitud o distancia más comunes utilizadas en la minería de datos se encuentran:

* Distancia Euclidiana: es una medida de distancia que se basa en la geometría euclidiana y mide la distancia en línea recta entre dos puntos en un espacio n-dimensional. Se utiliza comúnmente en problemas de clustering y clasificación.
* Distancia de Manhattan: también conocida como distancia de ciudad, se calcula sumando las diferencias absolutas entre las coordenadas de dos puntos. Esta distancia se utiliza con frecuencia en problemas de optimización.
* Correlación de Pearson: mide la similitud entre dos variables a través de su covarianza y varianza. Es una medida utilizada en problemas de análisis de correlación y asociación.
* Distancia de Hamming: se utiliza para medir la distancia entre dos cadenas de caracteres de igual longitud, y se cuenta el número de posiciones en las que los caracteres difieren.
* Distancia de Jaccard: mide la similitud entre dos conjuntos y se define como el cociente entre el número de elementos comunes en ambos conjuntos y el número total de elementos en ambos conjuntos.

Estas son solo algunas de las medidas de similitud más comunes utilizadas en la minería de datos, y cada una se utiliza para diferentes tipos de problemas y análisis. Es importante elegir la medida de similitud adecuada según el problema y los datos que se estén analizando.

**La distancia euclidiana:**

La distancia euclidiana es una medida de distancia comúnmente utilizada en problemas de minería de datos, aprendizaje automático y otros campos relacionados. Esta distancia se basa en la geometría euclidiana y mide la distancia entre dos puntos en un espacio n-dimensional.

La distancia euclidiana se calcula como la raíz cuadrada de la suma de las diferencias cuadráticas entre las coordenadas de dos puntos. Por ejemplo, si se tienen dos puntos en un espacio bidimensional (x1, y1) y (x2, y2), la distancia euclidiana entre ellos se calcula como:

distancia = sqrt((x2-x1)^2 + (y2-y1)^2)

Esta fórmula se puede generalizar para espacios n-dimensionales, donde la distancia euclidiana se calcula como:

distancia = sqrt((x2-x1)^2 + (y2-y1)^2 + ... + (zn - zn-1)^2)

La distancia euclidiana es una medida intuitiva y fácil de entender, ya que representa la distancia en línea recta entre dos puntos en un espacio n-dimensional. Se utiliza comúnmente en problemas de clustering y clasificación, donde se busca agrupar o clasificar objetos que se encuentran cerca en el espacio de características.

Sin embargo, es importante tener en cuenta que la distancia euclidiana puede ser sensible a la escala de los datos y a la presencia de características irrelevantes o redundantes. En algunos casos, puede ser necesario normalizar los datos antes de calcular la distancia euclidiana para evitar que las características con escalas más grandes dominen la medida de similitud.

En resumen, la distancia euclidiana es una medida de distancia comúnmente utilizada en la minería de datos y el aprendizaje automático. Es una medida intuitiva y fácil de entender que se utiliza para medir la similitud entre dos puntos en un espacio n-dimensional.

La distancia de Manhattan, también conocida como distancia rectilínea o distancia L1, es otra medida de distancia utilizada comúnmente en la minería de datos y el aprendizaje automático. Al igual que la distancia euclidiana, la distancia de Manhattan se utiliza para medir la similitud entre dos puntos en un espacio n-dimensional.

**La distancia de Manhattan:**

La distancia de Manhattan se calcula como la suma de las diferencias absolutas entre las coordenadas de dos puntos. Por ejemplo, si se tienen dos puntos en un espacio bidimensional (x1, y1) y (x2, y2), la distancia de Manhattan entre ellos se calcula como:

distancia = abs(x2 - x1) + abs(y2 - y1)

Esta fórmula se puede generalizar para espacios n-dimensionales, donde la distancia de Manhattan se calcula como:

distancia = abs(x2 - x1) + abs(y2 - y1) + ... + abs(zn - zn-1)

La distancia de Manhattan se llama así porque se asemeja a la forma en que se mide la distancia en una ciudad, donde uno puede moverse en línea recta solo en direcciones perpendiculares, como en una cuadrícula de calles.

La distancia de Manhattan se utiliza comúnmente en problemas de clustering, clasificación y búsqueda de vecinos más cercanos, donde se busca agrupar o clasificar objetos que tienen características similares. Al igual que la distancia euclidiana, es importante tener en cuenta que la distancia de Manhattan puede ser sensible a la escala de los datos y a la presencia de características irrelevantes o redundantes.

**Minería de Datos:**La minería de datos es una disciplina que utiliza técnicas y algoritmos de análisis de datos para extraer información útil y patrones interesantes a partir de conjuntos de datos grandes y complejos. Entre las técnicas de minería de datos más comunes se encuentran la clasificación, la regresión y el clustering.

* Clasificación: la clasificación es una técnica de minería de datos que se utiliza para categorizar o clasificar objetos en diferentes clases o categorías. En la clasificación supervisada, se utiliza un conjunto de datos de entrenamiento etiquetados para aprender un modelo que pueda clasificar nuevos objetos en las mismas clases. Los algoritmos de clasificación comunes incluyen el árbol de decisión, el k-NN (vecinos más cercanos), la regresión logística y las máquinas de soporte vectorial (SVM).
* Regresión: la regresión es otra técnica de minería de datos que se utiliza para predecir una variable continua en función de otras variables predictoras. Los algoritmos de regresión se utilizan comúnmente para analizar relaciones entre variables y para predecir valores futuros en un conjunto de datos. Los algoritmos de regresión comunes incluyen la regresión lineal, la regresión logística y la regresión polinómica.
* Clustering: el clustering es una técnica de minería de datos que se utiliza para identificar grupos o patrones en un conjunto de datos. En el clustering no supervisado, los datos no están etiquetados y el objetivo es encontrar patrones naturales o grupos dentro del conjunto de datos. Los algoritmos de clustering comunes incluyen el k-means, el clustering jerárquico, el clustering de densidad y el clustering espectral.

Además de estas técnicas, existen muchas otras técnicas de minería de datos, como la asociación de reglas, la detección de anomalías, el análisis de componentes principales (PCA), el análisis discriminante y la reducción de la dimensionalidad. Cada técnica tiene sus propias fortalezas y limitaciones, y es importante seleccionar la técnica adecuada para cada problema de minería de datos**.**

**Las redes neuronales:**

Las redes neuronales son un modelo de aprendizaje automático muy utilizado en la minería de datos debido a su capacidad para aprender y generalizar a partir de patrones en los datos. Estas redes consisten en capas de nodos interconectados que procesan los datos de entrada y producen una salida.

En la minería de datos, las redes neuronales se utilizan a menudo para tareas de clasificación y regresión, como la predicción de la probabilidad de que un cliente abandone una empresa o la predicción del precio de una vivienda. También se utilizan en el análisis de texto y en la visión por computadora, como en la clasificación de imágenes.

Las redes neuronales pueden ser entrenadas mediante el ajuste de los pesos de las conexiones entre los nodos, utilizando un conjunto de datos de entrenamiento etiquetado para aprender a identificar patrones y realizar predicciones precisas. Una vez entrenadas, las redes neuronales pueden ser utilizadas para hacer predicciones en nuevos datos que no fueron utilizados durante el entrenamiento.

**árboles de decisión en la minería de datos:**

Los árboles de decisión son un modelo de aprendizaje automático que se utiliza ampliamente en la minería de datos para realizar tareas de clasificación y regresión. Un árbol de decisión es una estructura en forma de árbol que representa un conjunto de reglas de decisión y permite la clasificación de nuevos datos en función de su similitud con los datos de entrenamiento.

En un árbol de decisión, cada nodo interno representa una pregunta sobre los datos y cada rama representa una posible respuesta. El proceso comienza en la raíz del árbol y se desplaza hacia abajo a través de los nodos internos, siguiendo el camino que se corresponde con las respuestas a las preguntas.

En la minería de datos, los árboles de decisión se utilizan a menudo para tareas de clasificación, como la identificación de patrones en los datos que permiten la clasificación de nuevos datos. Por ejemplo, un árbol de decisión podría ser utilizado para predecir si un cliente dejará o no una empresa en función de una serie de factores como la edad, la duración de la relación con la empresa, el historial de compras, etc.

Los árboles de decisión son fáciles de entender e interpretar, lo que los hace populares en muchos campos, incluyendo la medicina, la biología y la industria financiera. Además, pueden ser utilizados en conjunción con otros modelos de aprendizaje automático, como las redes neuronales y los métodos de clustering, para obtener resultados más precisos y completos.

**Naïve Bayes en la minería de datos:**

Naïve Bayes es un algoritmo de aprendizaje supervisado que se utiliza en la minería de datos para la clasificación y predicción de eventos. Se basa en el teorema de Bayes, el cual establece la probabilidad de que un evento ocurra en función de la probabilidad de que ocurran otros eventos relacionados.

En el contexto de la minería de datos, el algoritmo de Naïve Bayes se utiliza para clasificar elementos en diferentes categorías en función de las características que presentan. Se basa en la suposición de que las características de los elementos son independientes entre sí, lo que significa que la presencia o ausencia de una característica no afecta la presencia o ausencia de otras características.

El algoritmo de Naïve Bayes utiliza un conjunto de datos de entrenamiento para calcular la probabilidad de que un elemento pertenezca a una determinada categoría en función de sus características. A continuación, utiliza estas probabilidades para clasificar nuevos elementos que se presentan en función de sus características.

Naïve Bayes es especialmente útil en la minería de datos cuando el conjunto de datos es grande y las características de los elementos son independientes entre sí. También se utiliza con frecuencia en la clasificación de texto, como en la clasificación de correos electrónicos como spam o no spam.

**k-nearest neighbors :**

El algoritmo de k-vecinos más cercanos (k-nearest neighbors en inglés, abreviado k-NN) es un método de clasificación y regresión utilizado en la minería de datos y en el aprendizaje automático.

En el caso de la clasificación, el algoritmo k-NN determina la clase de un objeto desconocido basándose en las clases de los k objetos más cercanos a él en un conjunto de entrenamiento previamente clasificado. La elección del valor de k es importante en el rendimiento del algoritmo y puede determinarse mediante validación cruzada.

En el caso de la regresión, el algoritmo k-NN estima el valor de una variable continua para un objeto desconocido basándose en los valores de las k instancias más cercanas. Por ejemplo, se puede utilizar k-NN para predecir el precio de una casa en función de sus características como la ubicación, el número de habitaciones, el tamaño del lote, etc.

El algoritmo k-NN es simple de entender e implementar, y no requiere ningún modelo previo de entrenamiento. Sin embargo, puede ser computacionalmente costoso para grandes conjuntos de datos y no funciona bien en datos con muchas características o características irrelevantes.

**El clustering jerárquico:**

El clustering jerárquico es un método común de agrupamiento de datos en la minería de datos, que busca crear grupos homogéneos de datos en función de su similitud. El objetivo es clasificar los datos de manera que los elementos dentro de cada grupo sean lo más similares posible y los elementos entre grupos sean lo más diferentes posible.

En el clustering jerárquico, los datos se agrupan de manera jerárquica en una estructura de árbol. Se pueden distinguir dos tipos de clustering jerárquico: aglomerativo y divisivo.

En el clustering jerárquico aglomerativo, cada punto de datos comienza como su propio clúster, y los clústeres se van fusionando en grupos más grandes en función de la similitud entre ellos. En el clustering jerárquico divisivo, se comienza con todos los puntos de datos en un solo clúster y se dividen en clústeres más pequeños en función de la disimilitud entre ellos.

Los dendrogramas son una representación visual común del resultado del clustering jerárquico. Un dendrograma es un árbol en el que los nodos corresponden a los grupos de datos y las ramas corresponden a la similitud entre los grupos. La longitud de las ramas indica la distancia entre los grupos, y los grupos que están más cercanos en el árbol son los que tienen una mayor similitud.

El clustering jerárquico se puede utilizar en diversas aplicaciones, como segmentación de mercado, análisis de redes sociales, identificación de grupos en estudios biológicos, entre otras.