Taller Introducción a la Ciencia de Datos

Práctica: Explotando Conocimiento de los Datos

Impartido: MC. Héctor Alán de la Fuente Anaya
Email: hector.delafuente@cinvestav.mx

Horario: 19 Oct, 2022; 10:00 AM - 12:00 PM
Aula: Laboratorio de Cómputo

Ciencia de Datos: El objetivo de la ciencia de datos es encontrar conocimiento a partir del análisis de bases de datos. Se incluye cualquier técnica relacionada con el análisis de grandes colecciones de datos para descubrir patrones, tendencias, reglas, etc.

Objetivo: El Taller de Introducción a la Ciencia de Datos tiene por objetivo conocer la problemática que resuelve esta disciplina, dar una vista general de las técnicas más comunes en el preprocesamiento de los datos y llevar a cabo una práctica que permita descubrir conocimiento a partir de un conjunto de datos.

Requisitos

- **Herramienta de Ciencia de Datos:** Weka
  Enlace de descarga:
  `https://waikato.github.io/weka-wiki/downloading_weka`
- **Base de Datos**: Dataset `heart.csv`.
  Enlace de descarga:
  `https://drive.google.com/drive/folders/1JJofTu0cR1fXNqI3WY1xgSBkexv6bz7n`

Desarrollo

En esta práctica se extraerá conocimiento desde un conjunto de datos siguiendo el proceso metodológico *Descubrimiento de Conocimiento en Base de Datos* (KDD, por sus siglas en inglés), incluyendo cinco etapas: (1) Selección, (2) Preprocesamiento, (3) Transformación, (4) Minería de Datos e (5) Interpretación.

1. Selección

1. **Definir origen de los datos**
   En el modo *Explorador* de Weka, seleccionar Open file... para seleccionar el archivo de datos `heart.csv`.
2. **Identificar formato de los datos**
   Valores Separados por Comas (CSV, por sus siglas en inglés), la primera línea contiene los atributos.
3. **Identificar tipos de datos**
   Visualizar los datos por columnas y filas Edit..., identificar datos numéricos y datos categóricos.

2. Preprocesamiento

Limpieza de los datos

1. **Remover atributos seleccionando**
   Seleccionar Choose, filters > unsupervised > attribute > Remove. Configurar parámetros (índices de atributo).

2. **Remover valores mínimos**
   Seleccionar Choose, filters > unsupervised > instances > RemoveWithValues. Configurar parámetros (índices de atributo y punto de separación).

3. **Detectar outliers y valores extremos**
   Seleccionar Choose, filters > unsupervised > attribute > InterquartileRange. Configurar parámetros (todos los atributos).

4. **Remover outliers encontrados**
   Seleccionar Choose, filters > unsupervised > instances > RemoveWithValues. Configurar parámetros (índice de atributo outlier).

3. Transformación

1. **Normalización de datos:**
   Seleccionar Choose, filters > unsupervised > attribute > Normalize. Configurar parámetros (ignorar la clase).

4. Minería de Datos

Aprendizaje supervisado:

1. **Clasificación de datos:**
   Seleccionar Choose, classifiers > rules > ZeroR. Opciones de prueba (Seleccionar *Percentage split* %66).

5. Interpretación

1. **Evaluación de los resultados:**
   Identificar el porcentaje de instancias correctamente clasificadas en un set de prueba de %33.

2. **Mejorar los resultados de clasificación:**
   Tratar de mejorar el porcentaje de instancias correctamente clasificadas aplicando técnicas de preprocesamiento.

Actividad

Utilizando la base de datos *weather.arff*, y el clasificador ZeroR, tratar de mejorar el porcentaje de datos clasificados correctamente (60 %) mediante técnicas de preprocesamiento.

Enlace dataset weather

`http://www.cs.uni.edu/~jacobson/4772/Weka/datasets/weather.numeric.arff`