Entropy-SelApp

Manual Técnico

Tabla de contenido

1. Descripción

Actualmente, los conjuntos de datos que se utilizan en diferentes industrias en los procesos de ciencia de datos contienen millones de atributos, donde algunos atributos aportan un mayor grado de información al proceso de generación de conocimiento. Durante este proceso, toma lugar la etapa de preprocesamiento de datos, la cual se enfoca en manipular y transformar el conjunto de dato, haciendo que la información sea más entendible, coherente y productiva. El proceso de preprocesamiento de datos implica elegir una medida de resultado para evaluar, posibles variables influyentes, limpiar los datos y crear características El preprocesamiento de datos es un paso importante en el proceso de descubrimiento de conocimientos, porque las decisiones de calidad deben basarse en datos de calidad. Dentro de las principales tareas del preprocesamiento de datos, se encuentra la selección de atributo o características.

Un algoritmo de selección de atributos, es una solución computacional que tiene como finalidad determinar la relevancia de un atributo en el proceso de descubrimiento de conocimiento, a partir de una definición de una métrica la cual servirá como referencia para determinar dicha relevancia. En pocas palabras, estos algoritmos se centran en encontrar características relevantes y de esta manera mejorar el conjunto de datos utilizado en el descubrimiento de conocimiento. Una de estas técnicas es la selección de atributos basada en Entropía. La entropía como medida para ranquear características, se basa en eliminar N atributos, sin perder las características básicas del conjunto de datos, tomando como referencia una medida de similitud S que es inversamente proporcional a la distancia D entre dos instancias de N dimensiones. Estos cálculos se fundamentos en obtener la entropía del conjunto de datos eliminando cada uno de los atributos y calcular la diferencia frente a la entropía global del conjunto de datos y las entropías particulares obtenidas. Cuando todos los atributos son de tipo numérico, la medida de similitud S se obtiene mediante un método euclidiano, cuando todos los atributos son de tipo nominal se aplica el método de Hamming y en caso de tener un conjunto de datos heterogéneo es necesario aplicar métodos de discretización o numerización.

A partir de lo anterior, este producto de software realizar el proceso de selección de atributos mediante el método de entropía a un conjunto de datos seleccionado. De acuerdo a esto, permite cargar el conjunto de datos, automáticamente lo analizará y determinará cual distancia se debe utilizar si Euclideana o Hamming, o en caso de tener un conjunto de datos heterogéneo, realizará discretización de datos mediante la técnica de ChiMerge para posteriormente aplicar el método de Hamming. Al final de este proceso, el entorno expondrá al usuario cual atributo debería eliminarse de acuerdo a su valor de entropía.

2. Requisitos

Para el funcionamiento de este software, es necesario lo siguiente:

1. Python 3.8: https://www.python.org/downloads/release/python-380/
2. Flask 1.1.2: https://pypi.org/project/Flask/
3. NumPy 1.19 https://pypi.org/project/numpy/1.19.5/
4. Scipy 1.3.1 https://pypi.org/project/scipy/1.5.0/
5. Pandas 1.0.5 https://pypi.org/project/pandas/1.0.5/
6. Web Server for Chrome

3. Instalación

Para llevar a cabo la instalación, tenga en cuenta que es necesario que ya tenga instalado Python (mínimo la versión 3.8), de lo contrario, no será posible llevar a cabo los siguientes pasos. Cuando se instala Python, tendrá instalado PyPI (Python Package Index) por defecto, que es el que permite instalar paquetes de Python, con el comando pip y de esta forma realizar la instalación de librerías.

3.1. Clonación de repositorio

Una vez instalado Python en su máquina, clone el repositorio del proyecto de GitHub:

```
Git clone https://github.com/AndresRestrepoRodriguez/Entropy-SelApp.git
```

3.2. Instalación de librerías

Con el comando pip debe instalar las librerías de Python, de la siguiente forma:

```
pip install Flask == 1.1.2
```

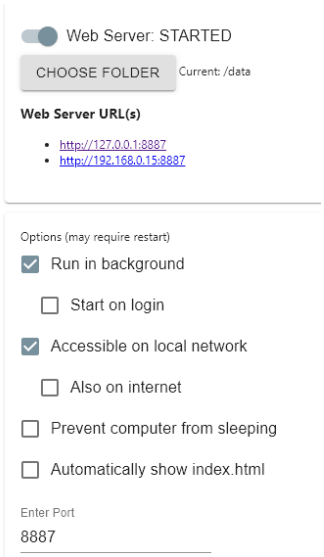
```
pip install numpy == 1.19.5
```

```
pip install scipy == 1.5.0
```

```
pip install pandas == 1.0.5
```

3.3. Web Server for Chrome

Si usted está trabajando en un entorno local, debemos tener un servidor web donde almacenar y alojar los distintos conjuntos de datos que se cargan a la plataforma para ser procesados, por esto es necesario instalar la extensión Web Server for Chrome. Una vez hemos instalado la extensión, se debe iniciar el servidor y seleccionar la carpeta que estará expuesta en la red, en este caso la ruta que se debe elegir será ***/ruta/usuario/Entropy-SelApp/static/data***. Adicionalmente, a esto, se presenta el puerto por donde se podrá acceder a dicha carpeta.



4. Ejecución de la aplicación

Para iniciar la aplicación, ubique su carpeta a través de la consola. Ahora, utilice el siguiente comando:

```
python core_app.py
```

El archivo core_app.py es el que le permitirá la activación del servidor; aguarde unos segundos mientras el proceso comienza, y una vez haya iniciado, se le indicará cuál es la dirección a la que deberá ingresar para utilizar la aplicación:

```
* Running on http://127.0.0.1:5000/
```