Data Warehousing & Data Mining LAB - G2
EXPERIMENT 2

- ASHISH KUMAR
- 2K18/SE/041

Aim:- List out various open source data mining tools and techniques and explain them.

Theory:-

Data Mining

Data mining is a process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Some of the most popular open-source tools used for Data Mining are:

1. Rapid Miner: RapidMiner is one of the best predictive analysis systems developed by the company with the same name as the Rapid Miner. It is written in JAVA programming language. It provides an integrated environment for deep learning, text mining, machine learning & predictive analysis. The tool can be used for over a vast range of applications including for business applications, commercial applications, training, education, research, application development, machine learning.
Rapid Miner offers the server as both on premise & in public/private cloud infrastructures. It has a client/server model as its base. Rapid Miner comes with template based frameworks that enable speedy delivery with reduced number of errors (which are quite commonly expected in manual code writing process).

2. Orange: Orange is a perfect software suite for machine learning & data mining. It best aids the data visualization and is a component based software. It includes a range of data visualization, exploration, preprocessing and modeling techniques and can be used as a module for the Python programming language. Orange has interactive data visualization and can also perform simple data analysis. As it is a component-based software, the components of orange are called 'widgets'. These widgets range from data visualization & pre-processing to an evaluation of algorithms and predictive modeling. Widgets offer major functionalities like:

- Showing data table and allowing to select features.
- Reading the data and training predictors and to compare learning algorithms.
- Visualizing data elements etc.

Additionally, Orange brings a more interactive and fun vibe to the dull analytic tools. Data coming to Orange gets quickly formatted to the desired pattern and it can be easily moved where needed by simply moving/flipping the widgets. Orange allows users to make smarter decisions in a short time by quickly comparing & analyzing the data.

3. Weka: Also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. It is best suited for data analysis and predictive modeling. It contains algorithms and visualization tools that support machine learning. Weka has a GUI that facilitates easy access to all its features. It is written in JAVA programming language and runs on almost any platform. Weka supports major data mining tasks including data mining, processing, visualization, regression etc. It works on the assumption that data is available in the form of a flat file. Weka can provide access to SQL Databases through database connectivity and can further process the data/results returned by the query.

4. KNIME: KNIME is the best integration platform for data analytics and reporting developed by KNIME.com AG. It operates on the concept of the modular data pipeline. It is a multi-language software development environment and comprises an integrated development environment (IDE) and an extensible plug-in system. KNIME constitutes of various machine learning and data mining components embedded together. KNIME has been used widely for pharmaceutical research. In addition, it performs excellently for customer data analysis, financial data analysis, and business intelligence. KNIME has some brilliant features like quick deployment and scaling efficiency. Users get familiar with KNIME in quite lesser time and it has made predictive analysis accessible to even naive users. KNIME utilizes the assembly of nodes to pre-process the data for analytics and visualization.

5. Apache Mahout: Apache Mahout is a project developed by Apache Foundation that serves the primary purpose of creating machine learning algorithms. It focuses mainly on data clustering, classification, and collaborative filtering. Mahout is written in JAVA and includes JAVA libraries to perform mathematical operations like linear algebra and statistics. Mahout is growing continuously as the algorithms implemented inside Apache Mahout are continuously growing. The algorithms of Mahout have implemented a level above Hadoop through mapping/reducing templates. To key up, Mahout has following major features

- Extensible programming environment.
- Pre-made algorithms and math experimentation environment.
- GPU computes for performance improvement.
- It allows applications to analyse large datasets in a faster manner.

6. DataMelt: DataMelt, also known as DMelt is a computation and visualization environment that provides an interactive framework to do data analysis and visualization. It is designed mainly for engineers, scientists & students. DMelt is written in JAVA and it is a multi-platform utility. It contains Scientific & mathematical libraries. Scientific libraries: To draw 2D/3D plots. Mathematical libraries: To generate random numbers, curve fitting, algorithms etc. Data Melt can be used for analysis of large data volumes, data mining, and stat analysis. It is widely used in the analysis of financial markets, natural sciences & engineering.

Some of the features are-

- DataMelt is a computational platform and can be used with different programming languages on various operating systems.

- DataMelt can be used with several scripting languages for the Java platform, such as Jython (Python programming language), Groovy, JRuby (Ruby programming language) and BeanShell.
- It creates high-quality vector-graphics images (SVG, EPS, PDF, etc.) that can be included in LaTeX and other text-processing systems.

Findings and Learning:

We learned about data mining and their respective significance. We also learned about the open source tools which can be used to assist us in the practice of data mining. Finally we learned about the features of the tools used for data mining.