Program – 2

AIM: List down the various open source data mining tools and techniques and their advantages and disadvantages

Introduction and Theory

Rapid Miner

Rapid Miner is one of the best predictive analysis system 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).

Rapid Miner constitutes of three modules, namely

1. Rapid Miner Studio- This module is for workflow design, prototyping, validation etc.
2. Rapid Miner Server- To operate predictive data models created in studio
3. Rapid Miner Radoop- Executes processes directly in Hadoop cluster to simplify predictive analysis.

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 has been written in Python computing language.

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

1. Showing data table and allowing to select features
2. Reading the data
3. Training predictors and to compare learning algorithms
4. Visualizing data elements etc.

Additionally, Orange brings a more interactive and fun vibe to the dull analytic tools. It is quite interesting to operate.

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. Users are quite fascinated by Orange. Orange allows users to make smarter decisions in short time by quickly comparing & analyzing the data.

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.

Program – 2

Weka has a GUI that facilitates easy access to all its features. It is written in JAVA programming language.

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.

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. 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.

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

1. Extensible programming environment
2. Pre-made algorithms
3. Math experimentation environment
4. GPU computes for performance improvement.

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 can run on any operating system which is compatible with JVM(Java Virtual Machine).

It contains Scientific & mathematical libraries.

Scientific libraries: To draw 2D/3D plots.

Mathematical libraries: To generate random numbers, curve fitting, algorithms etc.

DataMelt 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.

Program – 2

WEKA

Waikato Environment for Knowledge Analysis (Weka) is a suite of machine learning software written in Java, developed at the University of Waikato, New Zealand. It is free software licensed under the GNU General Public License.

Weka contains a collection of visualization tools and algorithms for data analysis and predictive modeling, together with graphical user interfaces for easy access to these functions. The original non-Java version of Weka was a Tcl/Tk front-end to (mostly third-party) modeling algorithms implemented in other programming languages, plus data preprocessing utilities in C, and a Makefile-based system for running machine learning experiments. This original version was primarily designed as a tool for analyzing data from agricultural domains, but the more recent fully Java-based version (Weka 3), for which development started in 1997, is now used in many different application areas, in particular for educational purposes and research.

Advantages of Weka include:

1. Free availability under the GNU General Public License.
2. Portability, since it is fully implemented in the Java programming language and thus runs on almost any modern computing platform.
3. A comprehensive collection of data preprocessing and modeling techniques.
4. Ease of use due to its graphical user interfaces.

Weka supports several standard data mining tasks, more specifically, data preprocessing, clustering, classification, regression, visualization, and feature selection. All of Weka's techniques are predicated on the assumption that the data is available as one flat file or relation, where each data point is described by a fixed number of attributes (normally, numeric or nominal attributes, but some other attribute types are also supported). Weka provides access to SQL databases using Java Database Connectivity and can process the result returned by a database query. Weka provides access to deep learning with Deeplearning4j. It is not capable of multi-relational data mining, but there is separate software for converting a collection of linked database tables into a single table that is suitable for processing using Weka. Another important area that is currently not covered by the algorithms included in the Weka distribution is sequence modeling.

Interface

Weka's main user interface is the Explorer, but essentially the same functionality can be accessed through the component-based Knowledge Flow interface and from the command line. There is also the Experimenter, which allows the systematic comparison of the predictive performance of Weka's machine learning algorithms on a collection of datasets.

The Explorer interface features several panels providing access to the main components of the workbench:

- The Preprocess panel has facilities for importing data from a database, a comma-separated values (CSV) file, etc., and for preprocessing this data using a so-called filtering algorithm. These filters can be used to transform the data (e.g., turning numeric attributes into discrete ones) and make it possible to delete instances and attributes according to specific criteria.
- The Classify panel enables applying classification and regression algorithms (indiscriminately called classifiers in Weka) to the resulting dataset, to estimate the accuracy of the resulting predictive model, and to visualize erroneous predictions,

receiver operating characteristic (ROC) curves, etc., or the model itself (if the model is amenable to visualization like, e.g., a decision tree).

- The Associate panel provides access to association rule learners that attempt to identify all important interrelationships between attributes in the data.
- The Cluster panel gives access to the clustering techniques in Weka, e.g., the simple k-means algorithm. There is also an implementation of the expectation maximization algorithm for learning a mixture of normal distributions.
- The Select attributes panel provides algorithms for identifying the most predictive attributes in a dataset.
- The Visualize panel shows a scatter plot matrix, where individual scatter plots can be selected and enlarged, and analyzed further using various selection operators.

Native Regression tools

Weka has a large number of regression and classification tools. Native packages are the ones included in the executable Weka software, while other non-native ones can be downloaded and used within R.Weka environment. Among the native packages, the most famous tool is the M5p model tree package.

Some of the regression tools are:

1. M5Rules (M5' algorithm presented in terms of mathematical function without a tree)
2. DecisionStump (same as M5' but with a single number output in each node)
3. M5P (splitting domain into successive binary regions and then fit linear models to each tree node)
4. RandomForest (several model trees combined)
5. RepTree (several model trees combined)
6. ZeroR (the average value of outputs)
7. DecisionRules (splits data into several regions based on a single independent variable and provides a single output value for each range)
8. LinearRegression
9. SMOreg (support vector regression)
10. SimpleLinearRegression (uses an intercept and only 1 input variable for multivariate data)
11. MultiLayerPerceptron (neural network)
12. GaussianProcesses.

There are several major data mining techniques have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree.

Association

Association is one of the best-known data mining technique. In association, a pattern is discovered based on a relationship between items in the same transaction. The association technique is used in market basket analysis to identify a set of products that customers frequently purchase together.

Retailers are using association technique to research customer’s buying habits. Based on historical sale data, retailers might find out that customers always buy crisps when they buy beers, and, therefore, they can put beers and crisps next to each other to save time for the customer and increase sales.

Classification

Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. Classification method makes use of mathematical techniques such as decision trees, linear programming, neural network, and statistics. In classification, we develop the software that can learn how to classify the data items into groups. For example, we can apply classification in the application that “given all records of employees who left the company, predict who will probably leave the company in a future period.” In this case, we divide the records of employees into two groups that named “leave” and “stay”. And then we can ask our data mining software to classify the employees into separate groups.

Regression

Regression, used primarily as a form of planning and modeling, is used to identify the likelihood of a certain variable, given the presence of other variables. For example, you could use it to project a certain price, based on other factors like availability, consumer demand, and competition. More specifically, regression’s main focus is to help you uncover the exact relationship between two (or more) variables in a given data set.

Clustering

Clustering is a data mining technique that makes a meaningful or useful cluster of objects which have similar characteristics using the automatic technique. The clustering technique defines the classes and puts objects in each class, while in the classification techniques, objects are assigned into predefined classes. To make the concept clearer, we can take book management in the library as an example. In a library, there is a wide range of books on various topics available. The challenge is how to keep those books in a way that readers can take several books on a particular topic without hassle. By using the clustering technique, we can keep books that have some kinds of similarities in one cluster or one shelf and label it with a meaningful name. If readers want to grab books in that topic, they would only have to go to that shelf instead of looking for the entire library.

Prediction

The prediction, as its name implied, is one of a data mining technique that discovers the relationship between independent variables and relationship between dependent and independent variables. For instance, the prediction analysis technique can be used in the sale to predict profit for the future if we consider the sale is an independent variable, profit could be a dependent variable. Then based on the historical sale and profit data, we can draw a fitted regression curve that is used for profit prediction.

Sequential Patterns

Sequential patterns analysis is one of data mining technique that seeks to discover or identify similar patterns, regular events or trends in transaction data over a business period.

In sales, with historical transaction data, businesses can identify a set of items that customers buy together different times in a year. Then businesses can use this information to recommend customers buy it with better deals based on their purchasing frequency in the past.

Decision trees

The A decision tree is one of the most commonly used data mining techniques because its model is easy to understand for users. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leads to a set of questions or conditions that help us determine the data so that we can make the final decision based on it.

Program – 2

Findings and Learnings:

1. There are wide variety of open source data mining tools available.
2. They vary greatly in their core specialties.
3. They also vary a lot in the way they approach data mining and the various algorithms they use.
4. Different techniques are available for data mining depending on the type of data and amount of data available.