EXPERIMENT: 10

(2K17/SE/79 PARV GUPTA)

AIM: Perform comparison of following data analysis tools.

a. WEKA
b. KEEL
c. SPSS
d. MATLAB
e. R

WEKA:

The Waikato Environment for Knowledge Analysis (Weka, http://www.cs.waikato.ac.nz/ml/Weka/) is a free and open source software package that assembles a wide range of data mining and model building algorithms. It does not support the creation of new features, though it does have support for automatic feature selection.

Weka has an extensive set of classification, clustering, and association mining algorithms that can be used in isolation or in combination, through methods such as bagging, boosting, and stacking. Users can invoke the data mining algorithms from the command line, a GUI (graphical user interface), or through a Java API. The command line interface and APIs are more powerful than the GUI, which does not give users access to all advanced functions. Weka can output the models it generates either in terms of the actual mathematical models, or in PMML (Predictive Modeling Markup Language) files which can be used to run the model on new data using the Weka scoring plugin to run the model.

Learning to use Weka is supported by a book by Witten, Frank & Hall (2011), now in its third edition. The Weka website also hosts an active mailing list, wiki, and bug reports.

KEEL:

KEEL (http://sci2s.ugr.es/keel/) is a data mining tool used by many EDM researchers. Unlike some of the tools listed above, which attempt to broadly survey different types of methods, KEEL has extensive support for some types of algorithms and tasks, but limited support for other algorithms and tasks. For instance, KEEL has extremely extensive support for discretization algorithms, but has limited support for other methods for engineering new features out of existing features. It has excellent support for feature selection, with a wider range of algorithms than any other package. It also has extensive support for imputation of missing data, and considerable support for data re-sampling.

For modeling, KEEL has an extensive set of classification and regression algorithms, with a large focus on evolutionary algorithms (although it is worth noting that evolutionary algorithms are currently not favored by many/most EDM researchers). Its support for other types of data mining algorithms, such as clustering and factor analysis, is more limited than other packages.

Support for association rule mining is decent, though not as extensive as some other packages. KEEL has relatively less support for new users than most other data mining packages, though there are help features and a user manual. KEEL is open-source and free for use under a GNU license.

SPSS:

Like Excel, SPSS is known beyond just the data science community. SPSS is primarily a statistical package, and offers a range of statistical tests, regression frameworks, correlations, and factor analyses. SPSS is complemented by IBM SPSS Modeler Premium, a relatively newer analytics and data mining package which integrates previous analytics and text mining packages. SPSS Modeler specifically has functionality for creating new features out of existing features, for data filtering, and for feature selection and feature space reduction. The tools for data transformation, feature selection, and feature space reduction are comparable to those seen in data mining packages, with a lower variety of selection approaches. There is also functionality for using the target class in feature selection, which is not available in many other packages. While SPSS represents a comprehensive statistical analysis tool, support for modeling is somewhat worse than the other tools in this section.

SPSS is less flexible than other tools, more difficult to customize, and is not documented as well. Support for procedures considered key by researchers in the educational data mining community, such as cross-validation, is also lacking when compared to tools more focused on data mining. SPSS is available commercially at http://www.ibm.com/analytics/us/en/technology/spss/.

MATLAB:

MATLAB is a programming language dedicated to mathematical and technical computing and it is designed for engineers and scientists. The desktop environment has a natural way of expressing computational mathematics such as linear algebra, data analytics, signal and image processing.

MATLAB features an application specific solution called ‘Toolboxes’. Toolboxes provide a set of MATLAB functions which are called as M-files that solves a specific set of problems. There are various areas where Toolboxes are available such as digital signal processing, control systems, neural network, simulations, Deep Learning, and many other areas. It has incomplete statistics support and is not open source. It has elegant matrix support and visualization. MATLAB, is a language that is easy to learn and remember because the syntax is simple and consistent by design across products, and hence MATLAB beats R.

R:

R is a popular and powerful open source programming language for statistical computing and graphics. R implements various statistical techniques like linear and non-linear modelling, machine learning algorithms, time series analysis, and classical statistical tests and so on. R consists of a language and a run-time environment with graphics, a debugger, access to certain system functions, and the ability to run programs stored in script files.

R is known for its steep learning curve. R was developed by statisticians hence its full capability is accessed through programming. There was no GUI to help non-programmers do the analysis. The working examples of R are complex and not for beginners. However, R-Commander and R-Studio the new GUI versions for R have benefitted the developer community.

R's main problem is its language, which, although highly extendable, is also a difficult one to learn thoroughly enough to become productive in DM. Advancement for DM tasks in that direction is the Rattle package that offers a decent GUI for R. Rattle, which is in development from 2006, is similar to Weka's Explorer in the sense of user-friendliness. It loads separate packages from R upon request for a specific analysis. Rattle uses some of the R's standard implementations of DM methods and also additional packages. The only problem with Rattleis that it cannot use all of the R's algorithms or Weka's DM implementations. Nevertheless, Rattle is user-friendly and quite popular in DM community.