International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

A Research Review on Comparative Analysis of Data Mining Tools and Techniques

Dr Paresh Ramnikbhai Rathod

Assistant Professor, Shree Swaminarayan College of Computer Science, Bhavnagar, Gujarat

Abstract

In today's world, data mining is essential to both the corporate sector and society at large. Numerous proprietary and open-source tools for data mining and visualization are available to help glean insightful information from enormous data stores. Data mining is a crucial procedure in many domains, such as business, healthcare, education, and scientific research, because it is an effective method for extracting insightful information from large and complicated datasets. In many domains, data mining the process of drawing insightful conclusions from massive datasets has become an essential tool. Researchers and practitioners can successfully use data mining techniques to tackle difficult problems and obtain a competitive edge by being aware of these tools' potential. The Environment for Knowledge Analysis (WEKA), Konstanz Information Miner (KNIME), GhostMiner, R Analytical Tool to Learn Easily (Rattle), and RapidMiner are among the several data mining technologies reviewed in this research, and are provided under the GNU GPL licenses while GhostMiner is commercial. Most of the time, selecting a data mining tool to do research is a barrier for novice researchers. This report also includes an assessment of the sources, qualities, and capabilities.

Keywords

Data Mining, RapidMiner, Weka, Knime, GhostMiner, Rattle, Data Exploration, Data Modelling

1. Introduction

Information technology has become increasingly significant in both the social and business spheres in the late 20th and early 21st centuries. Therefore, among other things, decision assistance, business performance management, query and reporting analytical processing, and predictive analysis all require various facets of information technology. Data mining is one field of information technology that has become increasingly significant in the modern world for predictive analysis and decision assistance. Studies and research show that data mining is crucial in assisting companies in analyzing data and, consequently, in making defensible choices about many facets of their operations and procedures. Finding patterns, trends, and useful insights in big datasets requires the use of data mining techniques. They make processes like applying machine learning algorithms, pre-processing data, and visualizing outcomes easier. These technologies assist professionals derive meaning from complicated data structures, which is crucial in fields like scientific research, business analytics, and healthcare. From this vantage point, several data mining technologies will be assessed in this study. These consist of

Rapid Miner, R Analytical Tool To Learn Easily (Rattle), GhostMiner, Konstanz Information Miner (KNIME), and Waikato Environment for Knowledge Analysis (WEKA).

REVIEW OF DATA MINING TOOLS

1. **Waikato Environment for Knowledge Analysis (WEKA)**
2. **Konstanz Information Miner (KNIME)**
3. **GhostMiner**
4. **R Analytical Tool to Learn Easily (Rattle)**
5. **RapidMiner**

1. Waikato Environment for Knowledge Analysis (WEKA)

One of the most well-known machine language and data mining programs is WEKA. WEKA was founded in 1992 with financial support from the government of New Zealand. This tool, which uses the Java programming language, is well known for its data mining features. Data pre-processing is one of WEKA's primary functions. In light of this, WEKA can convert unprocessed data into a comprehensible format. claims that this data mining tool offers a large selection of data pre-processing tools or instead of filters that let people do various things with data. These tools/filters include: Random subset, RELAGGS, reservoir sample, subset by expression, kernel filter, numerical cleaner, numerical to nominal, propositional to multi-instance and vice versa, add classification, add ID, add values, attribute reorder, interquartile range, and wavelet. The software also has the ability to classify data. The software employs four primary phases for data classification in order to do this: data preparation, algorithm selection and application, tree generation, and output or result analysis. Furthermore, WEKA can cluster data using unsupervised methods like this. WEKA clusters data using a variety of techniques, such as density-based clustering, hierarchical clustering, and K-means clustering. Likewise, this program accepts a number of data files, such as those in the ARFF, CSV, LibSVM, and C4.5 formats, and WEKA 3.6 allows for the import of PMML regression. Additionally, it can read information from files, URLs, and databases that support the JDBC driver. The Explorer, Experimenter, Knowledge Flow, and Simple CLI are the four primary components of the WEKA data mining software's user interface. The extensibility features of WEKA are also implemented through the use of three plugins. Importantly, WEKA is an open source, is distributed freely under the GNU General Public License, and if fully portable.

2. Konstanz Information Miner (KNIME)

Based on Eclipse and developed in Java, Konstanz Information Miner (KNIME) is an open source platform for data mining, analytics, integration, and reporting. The GPLv3 license governs the release and provision of KNIME. There is a commercial edition of KNIME, better suited for big corporations and organizations, for users that want specialized support. However, the majority of users would still find this software's free edition helpful for their data mining tasks. Crucially, the program features a graphical user interface that makes it simple and rapid for users to put up nodes for data processing extraction, transformation, and leading, or ETL so that they may model, pre- process, and display data. The data loading, processing, analysis, transformation, and visual exploration modules that KNIME offers don't have to focus on any one aspect of the application. KNIME's primary features include regular

upgrades, adaptability, and versatility. claims that KNIME demonstrates its versatility by working with data created in different programs like Python, R, and Java. Although this is true, it is important to remember that KNIME can only read data from ARFF files, which include databases, text-delimited files, WEKA files, and CSV files. Data from Microsoft Excel files, LIBSVM files, and SVMlight files cannot be read by KNIME. It also doesn't make use of SPSS files. From this angle, people who are already familiar with the LIBSVM and SVMlight file formats will find this software challenging to use.

3. GhostMiner

Fujitsu created the data mining program GhostMiner, which is sold commercially. GhostMiner supports a wide range of data and data formats, including databases (including spreadsheets) and sophisticated machine learning algorithms. The program is essential for data preparation and selection, visualization, and model and multimodel validation. This software has several important features. First off, GhostMiner can handle data from a variety of file formats, such as text, ASCII files, and spreadsheets. Crucially, the program contains features that let customers create basic user interfaces based on their data mining requirements. Likewise, data visualization and data pre-processing are among its primary functions. The majority of statistical operations in preliminary statistical data analysis, including determining the mean and median over the entire data on text, ASCII, and database files, are made possible by GhostMiner's data pre-processing capabilities. In terms of visualization, GhostMiner may demonstrate the relationships between a set or sets of data. Along with these features, this data mining program offers a number of statistical tools, data modeling tools, and techniques like Xtest and cross-validation to assess the precision of the data being studied. However, compared to comparable technologies from firms like IBM and Oracle, this tool offers minimal support and is limited when dealing with complex data mining processes. As a result, small and midsize users, not huge corporations, are advised to utilize the program.

4. R Analytical Tool To Learn Easily (Rattle)

Rattle is a commercial, open source data mining product that is developed in the R programming language and offers a link into R. Rattle makes use of a Gnome graphical user interface, which is implemented by the RGtk2 package per. Crucially, the graphical user interface of this software was made with the interactive interface builder Glade; as a result, it is not dependent on the XML description. Rattle can connect to a variety of data sources, including MySQL, Oracle, Teradata, SQLite, SQL Server, IBM DB2, Microsoft Access, Excel, Postgress, and others, and supports a number of data types, including CSV, TXT, ARFF, R datasets, and ODBC databases.
It's also important to remember that Rattle offers two sophisticated tools that facilitate interactive graphical data analysis: Latticist and GGobi. On the one hand, the GGobi tool's highly dynamic and interactive graphics, including bar charts, tours, parallel coordinate plots, and scatterplots, are useful for examining high-dimensional data. Crucially, GGobi supports Windows, OS/X, and Linux platforms and must be installed independently on the systems.
The Latticist software, on the other hand, analyzes data using a graphical user interface with lattice visuals. Among other things, the tools include data selection, annotations, plots, brushing, and subsetting.

Furthermore, the Rattle may choose, explore, visualize, convert, cluster, associate, model, and evaluate data, among other processes.

5. RapidMiner

One data mining tool from the Data Mining Suites (DMS) is RapidMiner. RapidMiner is an open source program that is published under the GNU Affero General Public License (AGPL), in contrast to the majority of DMS data solutions, which are commercial. RapidMiner is a Java-based tool that may be used for text, data, and web mining. There are several features in RapidMiner. Several user-friendly interfaces, simple data management and accessibility, model import and export, an operator for applying models to data sets known as scoring, data partitioning, data replacement, graphs and visualization, attribute generation, and Bayesian modeling are just a few of the features of this data mining software, according to reference. In a similar vein, observes that tokenization, extraction, stemming, transformation, utility, and other operations including document creation, writing, reading, and processing are a few of the text mining operators. RapidMiner is capable of parsing and mining text in this instance. Regarding, RapidMiner offers over 400 operators for Java data mining. Furthermore, this software has drag-and-drop functionality, an intuitive graphical user interface (GUI), and features an application wizard with choices to help process data automatically based on project goals. This software, for instance, can assist in processing data based on sentiment analysis, direct marketing, and churn analysis. Crucially, RapidMiner has several commercial versions in addition to being an open source data mining program.

Installing RapidMiner

- **Windows**: Download and install using rapidminer-install.exe. By default, it installs under C:\Program Files and is accessible via the Start menu.
- **macOS**: Download the .dmg file and add it to the Applications folder.

RapidMiner Products

A range of products are available from RapidMiner to meet different machine learning and data analysis requirements. These packages offer all-inclusive tools for tasks including deployment, model construction, and data preparation. Some of the main items in the RapidMiner suite are listed below:

1. **RapidMiner Studio**
   - **Overview**: A strong instrument for analyzing non-structured data, including multimedia, text, and photos. It effectively analyzes unstructured data by converting it into structured formats.
   - **Features**:
     - Data loading, manipulation, and access.
     - Support for both sophisticated and conventional data modeling methods.
     - A user-friendly interface for creating data pipelines.
2. **RapidMiner Auto Model**

- **Overview**: Building and validating data models is made easier with Auto Model, an enhanced version of RapidMiner Studio.
- **Capabilities**:
  - Focuses on outlier detection, grouping, and predictive modeling.
  - Allows for process customisation to satisfy particular needs.
  - simplifies processes for implementation in practical applications.

3. **RapidMiner Turbo Prep**

- **Overview**: Turbo Prep was created especially to handle the time-consuming nature of data preparation, guaranteeing a smooth and effective procedure.
- **Features**:
  - Throughout the preparation process, the data is visible thanks to an intuitive interface.
  - enables users to make small adjustments and observe outcomes right away.
  - provides a wide range of operations for transforming data so that it is prepared for presentation or model construction.

These products collectively enhance the data science lifecycle by offering specialized tools for each stage, from data pre-processing to advanced analytics and deployment.

Features of RapidMiner

1. **User-Friendly Application & Interface**

- RapidMiner Studio is a visual data science workflow designer that facilitates model validation and prototyping, enabling users to work more productively.

2. **Versatile Data Access**

- Able to analyze both organized and unstructured data, including multimedia, text, and images.
- It can convert unstructured data into formats that are structured, making it compatible with a range of analysis tools.

3. **Effective Data Exploration**

- RapidMiner helps users quickly create plans for data preparation by providing tools for rapid data exploration.
- Key statistics and insights are automatically extracted for a thorough comprehension of the material.

4. **Advanced Data Preparation**

- Outfitted with a wide range of tools to tackle intricate data transformation problems, allowing users to best prepare data for predictive analysis.
- Allows users to take advantage of the entire range of data accessible for analysis by integrating both structured and unstructured data.
- Workflows for data preparation can be stored for later usage, increasing efficiency.
-

5. **Powerful Modeling Capabilities**

- A wide range of machine learning methods for supervised and unsupervised learning are available in RapidMiner Studio.
- Provides strong, adaptable tools for creating models that are suited to any particular commercial or research requirement.

6. **Accurate Model Validation**
   - To guarantee accurate and reliable model performance estimations, a modular validation technique is employed.
   - The platform minimizes overfitting and ensures the integrity of the findings by preventing information from leaking between preprocessing and model training.
7. **Streamlined Scoring**
   - Makes it simple to apply the model, either for usage inside the RapidMiner platform or for incorporating the output into other systems.
8. **Enhanced Process Control & Management**
   - Designed for intricate use cases, RapidMiner Studio streamlines the process for non-programmers by requiring no coding knowledge.
   - Provides sophisticated process control tools that let users create repeatable procedures, automate operations, and effectively manage workflows.
   - Compatible with a number of scripting languages for users who want to increase its functionality.

Example Sets and Attributes

Once data is loaded into RapidMiner, it is stored in an **example set**. An example set is a collection of data instances, each described by a set of **attributes** (also known as features). These attributes hold the values that will be processed during the mining process.

Each attribute has:

- **Value Types**: These define the nature of the data and how it will be processed:
  - **Numeric Data**: Data that has a natural order, meaning values have a meaningful scale or distance between them. For example, in numeric data, 2 is closer to 1 than it is to 5.
  - **Nominal Data**: Categorical data that has no inherent order. For example, the colors red, green, and blue are nominal attributes, meaning they are distinct categories, but there's no order or ranking between them.
- **Roles**: Every attribute is also assigned a role, which determines how the data will be treated in the analysis:
  - **Regular**: Attributes that contain the input data for analysis.
  - **Label**: The target attribute for tasks like classification or regression (the output variable).
  - **Id**: A unique identifier for each example in the set.
  - **Weight**: The importance or frequency of an example in the analysis.

By designing a process with these operators and understanding the types and roles of attributes, users can manipulate and analyze data more effectively, creating tailored workflows to suit various data mining tasks.

Table 1: Different Value Type

Value Type	Description
binominal	Only two different values are permitted
polynomial	More than two different values are permitted
integer	Whole numbers, positive and negative
real	Real numbers, positive and negative
date_time	Date as well as time date Only date-time Only time

Table 2: Types of Roles

Role	Description
Id	A unique identifier, no two examples in an example set can have the same value
Regular (default)	Regular attribute that contains data
Label	The target attribute for classification tasks
Weight	The weight of the Examples concerning the label
Cluster	Created by RapidMiner as the result of a clustering task
Prediction	Created by RapidMiner as the result of a classification task

How to Use RapidMiner

Using RapidMiner is a straightforward process thanks to its intuitive interface. Below are the key steps to help you get started with creating a data mining process using the **Design Perspective**.

The **Design Perspective** is where you create and manage your data mining processes. Here are the core elements:

- **View Your Current Process**: The "Process" tab shows the flow of your data mining pipeline. It provides a visual representation of all the steps, from data loading to model building and evaluation.
- **Access Your Data and Processes**: The **"Repository"** panel allows you to access and organize your datasets, models, and other components needed for your analysis. You can easily load data from the repository and use it in your processes.
- **Add Operators to the Process**: Operators are the building blocks of a RapidMiner process. They are used to perform various tasks such as data loading, data transformation, and model training. You can drag and drop operators from the **"Operators"** panel onto your workflow. Operators represent specific functions like data cleaning, classification, or clustering.
- **Configure the Operators**: After adding an operator, you need to configure it to fit your needs. This is done through the **"Parameters"** panel. Here, you can adjust settings such as the type of algorithm to use, the parameters for training a model, or the method of data preprocessing.
- **Learn About Operators**: If you are unsure about how to use a particular operator, the **"Help"** section provides detailed information and tutorials on the operator's functionality and options. This is useful for understanding the purpose and configuration of each operator.

Build Your Workflow

Once you have a basic understanding of how to add and configure operators, you can start building a workflow that suits your data analysis needs:

- **Data Preparation**: Begin by adding data access operators (e.g., Retrieve, Read) to load your dataset. You can also perform data cleansing and transformation using operators like Filter or Normalize.
- **Modeling**: Use machine learning operators (e.g., Decision Tree, K-Means, Regression) to build predictive models.
- **Evaluation**: Evaluate the model performance using operators such as Cross Validation or Performance to test the model on a different dataset.

Run the Process

Once you have designed your workflow and configured all operators, click the **Run** button to execute your process. RapidMiner will process the data and provide you with the output results, which can include statistical summaries, visualizations, or model predictions.

Inspect the Results

After running your process, use the **Results Perspective** to inspect the output. The **Data View**, **Statistics View**, and **Visualizations View** will help you analyze the results in detail.

By following these steps, you can effectively use RapidMiner to create, configure, and execute data mining processes, making the most of the powerful analytics tools it provides.

Using the "Results Perspective" in RapidMiner

The **Results Perspective** in RapidMiner is where you can view and analyze the results of the processes you've run. This view provides various panels to help you inspect the output and gain insights from your data. Here's a breakdown of the key features of the **Results Perspective**:

1. Data View

- The **Data View** displays the **ExampleSet** (your dataset) that was generated by your process. This view lets you examine the data output from your operators.
- You can see each attribute or feature of your dataset and their corresponding values. This is especially helpful for checking if your data has been transformed or processed as expected.
- You can filter, sort, or search through the data to analyze specific examples or attributes. This is useful if you're looking for patterns or anomalies in your output data.

2. Statistics View

- The **Statistics View** provides detailed information about your dataset's metadata and statistical properties.

- It includes key metrics such as **mean**, **standard deviation**, **min**, **max**, and other relevant statistics for each attribute in the dataset.
- This is helpful for understanding the distribution and variability of your data, which can guide further analysis or data transformations.
- If you are working with classification or regression models, this view may also show the performance metrics such as accuracy, error rate, or AUC (Area Under the Curve) for the model.

3. Visualizations View

- The **Visualizations View** allows you to generate graphical representations of your data. This is a powerful feature for exploring trends, correlations, and distributions visually.
- You can create various types of visualizations, such as:
  - **Histograms**: To visualize the distribution of data values across attributes.
  - **Scatter Plots**: To examine relationships between two or more variables.
  - **Box Plots**: To detect outliers and understand the range of your data.
  - **Confusion Matrices**: If you are working with classification models, this visualization helps assess how well your model predicted different classes.

By using these views, you can thoroughly analyze your results, make informed decisions based on data trends, and gain a deeper understanding of how your models or processes are performing. Each view in the **Results Perspective** offers a different angle on your data, making it easier to uncover insights and refine your analysis.

Summary of strengths and weaknesses of the data mining tools.

Data mining tool	Strengths	Weaknesses
WEKA	Robust data mining techniques, Support web services	Lack of colourful visualization
KNIME	Adaptability, No need for programming knowledge	Can only support ARFF file format, No support for web services
GhostMiner	Wide range of statistical and data modeling techniques	Cannot not handle complex data, Only for small or medium business
Rattle	Active user Group	Not meant for novice or learners
Rapid Miner	User friendly, Application wizard, High level visualization support, Support web services	Need for programming knowledge to use to effectively use the software

CONCLUSION

RapidMiner is powerful approaches to data mining. RapidMiner excels in versatility and user-friendly automation, making it ideal for businesses and advanced analytics projects. It provides robust machine learning algorithms for academic and experimental use, focusing on structured datasets. Choosing the right tool depends on the specific needs of the project, user expertise, and desired outcomes.

To sum up, data mining is essential to the contemporary business environment's decision-making process. Businesses can process and analyze data using a variety of data mining software programs. These consist of WEKA, KNIME, GhostMiner, RapidMiner, and Rattle. It might be argued that GhostMiner is a commercial program that can only be purchased, but Rattle, KNIME, RapidMiner, and WEKA are open source programs that are freely available under the GNU GLP licenses. Furthermore, although Rattle, KNIME, RapidMiner, and WEKA are capable of high-level data mining, functions, GhostMiner is limited to low level data mining functions

References

1. Data ware housing and Data Mining **– By Harjinder Kaur**
2. Data Mining and Data Warehousing: Principles and Practical Techniques – **By Parteek Bhatia**
3. Data Mining: Concepts and Techniques - By **Jiawei Han, Micheline Kamber, Jian Pei**
4. Introduction to Data Mining **– by Pang-Ning Tan, Michael Steinbach, Vipin Kumar**
5. Barcaroli G., Bergamasco S., Jouvenal M., Pieraccini G., Tininini L. Generalised software for statistical cooperation, 2008, [Online]. Available from:www.researchgate.net/profile/Giulio_Barcaroli/publication/263923 216_Generalised_software_for_statistical_cooperation/links/0f31753c52fc97f19d000000 .pdf [Accessed 29 March 2015].Akhtar, F. and Hahne, C. Rapid Miner 5 Operator Reference. Rapid-I GmbH, 2012, Retrieved 13:15, February 13, 2015 from: http://rapidminer.com/wpcontent/uploads/2013/10/RapidMiner_OperatorReference_en.pdf.
6. Bulusu, L. Open source data warehousing and business intelligence. Boca Raton, FL: CRC Press, 2012. Peterson, Leif E. "K-nearest neighbor." Scholarpedia 4(2), 1883, 2009.
7. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., Witten I. H. ‘The WEKA data mining software: an update’. ACM SIGKDD Explorations Newsletter, 11(1), 10–18, 2009.
8. Joseph, M. V. ‘Significance of data warehousing and data mining in business applications’. International Journal of Soft Computing and Engineering, 3(1), 329–333, 2013.
9. Jovic, A., Brkic, K. and Bogunovic, N. ‘An overview of free software tools for general data mining’. Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2014 37th International Convention, 1112–1117, 2014.
10. Lin, N. ‘Applied business analytics: Integrating business process, big data, and advanced analytics’. New Jersey: Pearson Education Ltd., 2014.
11. Melin, P., Kacprzyk, J. and Pedrycz, W.’Soft computing for recognition based on biometrics’. Berlin Heidelberg: Springer Science & Business Media, 2010.
12. Pierre, T. F. Software applications: Concepts, methodologies, tools, and applications: concepts, methodologies, tools, and applications. Hershey, PA: IGI Global, 2009.
13. Rapid-I GmbH. (n.d.) Fact Sheet: RapidMiner and RapidAnalytics – Business analytics fast and powerful [Online]. Available from: http://www.rapidi.com/downloads/brochures/RapidMiner_Fact_She et.pdf [Accessed 29 March 2015].
14. Shterev, J. ‘Demo: Using RapidMiner for Text Mining’. Digital Presentation and Preservation of Cultural and Scientific Heritage, 3, 254–256, 2013.

15. Singhal, S. and Jena, M. ‘A study on WEKA Tool for data pre-processing, classification and clustering’. International Journal of Innovative Technology and Exploring Engineering (IJITEE), 2 (6), 250–253, 2013.
16. Wang, J., Hu, X., Hollister, K. and Zhu, D. ‘A comparison and scenario analysis of leading data mining software’. in Selected readings on information technology management: Contemporary issues. ed. by Kelley, G. Hershey, PA: IGI Global, 2008.
17. Williams, G. J. ’Rattle: A data mining GUI for R.’ The R Journal, 1(2), 45–55, 2009.