ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

A Study on Adoption of Data Mining Tools and Collision of Predictive Techniques

Lavanya.M[1], Dr.B.Jagadhesan[2]

M.Phil Research Scholar, Dhanraj Baid Jain College (Autonomous), Thoraipakkam, Chennai, India[1]

Associate Professor, Dhanraj Baid Jain College (Autonomous), Thoraipakkam, Chennai, India[2]

ABSTRACT: Data mining, also known as knowledge discovery from databases, is a process of mining and analysing enormous amounts of data and extracting information from it. Data mining can quickly answer business questions that would have otherwise consumed a lot of time. Some of its applications include market segmentation – like identifying characteristics of a customer buying a certain product from a certain brand, fraud detection – identifying transaction patterns that could probably result in an online fraud, and market based and trend analysis – what products or services are always purchased together, etc. This article focuses on the various open source options available and their significance in different contexts.

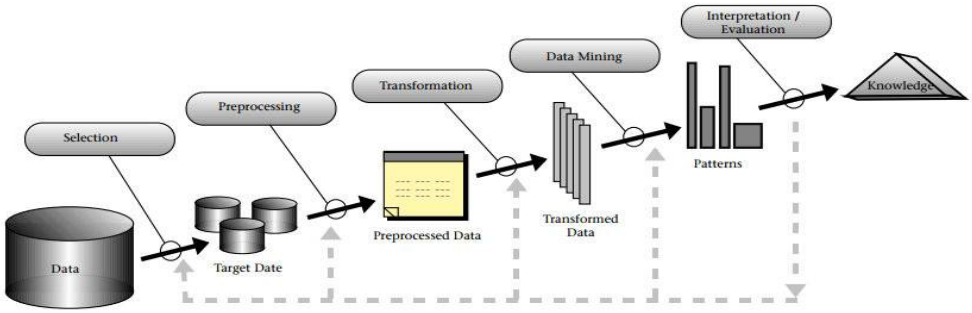
KEYWORDS: Data Mining, Clustering, Mining Tools.

I. INTRODUCTION

The development and application of data mining algorithms requires the use of powerful software tools. As the number of available tools continues to grow, the choice of the most suitable tool becomes increasingly difficult. Some of the common mining tasks are as follows.

- Data Cleaning − The noise and inconsistent data is removed.
- Data Integration − Multiple data sources are combined.
- Data Selection − Data relevant to the analysis task are retrieved from the database.
- Data Transformation − Data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- Data Mining − Intelligent methods are applied in order to extract data patterns.
- Pattern Evaluation − Data patterns are evaluated.
- Knowledge Presentation − knowledge is represented.

The following diagram shows the process of knowledge discovery:



ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

Pre-processing: This involves all the preliminary tasks that can help in getting started with any of the actual mining tasks. Pre-processing could be removing anomalies and noise from the data that's about to be mined, filling in missing values, normalising the data or compressing data using techniques like generalisation and aggregation.
Clustering: This is partitioning a huge set of data into related sub-classes.
Classification: This is tagging or classifying data items into different user-defined categories.
Outlier Analysis: helps in identifying those data elements which are deviant or distant from the rest of the elements in a dataset. This can help in anomaly detection.
Associative Analysis: It helps in bringing out hidden relationships among data items in a large data set. This can help in predicting the occurrence of a particular item in a transaction or an event whenever some other item is present. You can think of this as a conditional probability.
Regression: It is used to predict values of a dependent variable by constructing a model or a mathematical function out of independent variables.
Summarisation: It helps in coming up with a compact description for the whole data set. Data mining is a combination of various techniques like pattern recognition, statistics, machine learning, etc. While there is a good amount of intersection between machine learning and data mining, as both go hand in hand and machine learning algorithms are used for mining data, we will restrict ourselves in this article to only those tools specialised for data mining.

II.IMPACT OF DATA MINING TOOLS

Weka: It is a Java based free and open source software licensed under the GNU GPL and available for use on Linux, Mac OS X and Windows. It comprises a collection of machine learning algorithms for data mining. It packages tools for data pre-processing, classification, regression, clustering, association rules and visualisation. The various ways of accessing it are – Weka Knowledge Explorer, Experimenter, Knowledge Flow and a simple CL. Explorer is a user-friendly graphical interface for two-dimensional visualisation of mined data. It lets you import the raw data from various file formats, and supports well known algorithms for different mining actions like filtering, clustering, classification and attribute selection. However, when dealing with large data sets, it is best to use a CL based approach as Explorer tries to load the whole data set into the main memory, causing performance issues. This software also provides a Java Appetiser for use in applications and can connect to databases using CJD. Weka has proved to be an ideal choice for educational and research purposes, as well as for rapid prototyping.
Rapid Miner: It is available in both FOSS and commercial editions and is a leading predictive analytic platform. Gartner, the US research and advisory firm, has recognised Rapid Miner and Knife as leaders in the magic quadrant for advanced analytic platforms in 2016. Rapid Miner is helping enterprises embed predictive analysis in their business processes with its user friendly, rich library of data science and machine learning algorithms through its all-in-one programming environments like Rapid Miner Studio. Besides the standard data mining features like data cleansing, filtering, clustering, etc, the software also features built-in templates, repeatable work flows, a professional visualisation environment, and seamless integration with languages like Python and R into work flows that aid in rapid prototyping. The tool is also compatible with weak scripts. Rapid Miner is used for business/commercial applications, research and education.
Orange: Python users playing around with data sciences might be familiar with Orange. It is a Python library that powers Python scripts with its rich compilation of mining and machine learning algorithms for data pre-processing, classification, modelling, regression, clustering and other miscellaneous functions. Orange also comes with a visual programming environment and its workbench consists of tools for importing data, and dragging and dropping widgets and links to connect different widgets for completing the workflow. The visual programming comes with an easy-to-use UI, with plenty of online tutorials for assistance. Due to the ease of programming and integration in Python, Orange can be a great take off point for novices and experts to plunge into data mining.
Knime:is one of the leading open source analytic, integration and reporting platforms that comes as free software and as well as a commercial version. Written in Java and built upon Eclipse, its access is through a GUI that provides options to create the data flow and conduct data pre-processing, collection, analysis, modelling and reporting. A Gartner survey reveals that customers are happy with the platform's flexibility, openness and smooth integration with other software like Weka and R. Given the small size of the company, Knime has a large user base and an active community. It makes use of Eclipse's extension mechanism capability to add plugins for the required functionalities like text and image mining. This software is ideal for enterprise use.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

DataMelt or DMelt does much more than just data mining. It is a computational platform, offering statistics, numeric and symbolic computations, scientific visualisation, etc. To avoid digressing from our topic, I'll restrict myself to only covering its data mining capabilities. DMelt provides data mining features like linear regression, curve fitting, cluster analysis, neural networks, fuzzy algorithms, analytic calculations and interactive visualisations using 2D/3D plots and histograms. One can play around with its IDE (integrated development kit) or its functions can be called from applications using its Java API. Both community and commercial editions of DMelt are available on Linux, Mac OS, Windows and Android platforms. DMelt is a successor to the jHepWork and SCaVis programs, which some people working in data analysis might be familiar with. This software is well suited for students, engineers and scientists.

Apache Mahout: is primarily a library of machine learning algorithms that can help in clustering, classification and frequent pattern mining. It can be used in a distributed mode that helps easy integration with Hadoop. Mahout is currently being used by some of the giants in the tech industry like Adobe, AOL, Drupal and Twitter, and it has also made an impact in research and academics. It can be a great choice for anyone looking for easy integration with Hadoop and to mine huge volumes of data.

ELKI: It is open source software written in Java and licensed under AGPLv3. This software focuses especially on cluster analysis and outlier detection with a compilation of numerous algorithms from both these domains. The software is accessed through a GUI that displays the results once the selected algorithm is run. ELKI's design goals are performance, scalability, completeness, extensibility and a modular design to welcome contributions. ELKI currently doesn't offer professional support and the software is optimised for use in science and research. Hence, this option works best for those in research.

MOA: Massive Online Analysis (MOA), as the name suggests, is primarily data stream mining software that is well suited for applications that need to handle volumes of real-time data streams at a high speed. MOA is distributed under GNU GPL, and can be used via the command line, GUI or Java API. It is a rich compilation of machine learning algorithms and has proved to be a great choice during the design of real-time applications. Stream mining algorithms typically require faster computations without storing all of the datasets in the memory and have to get the work done within a limited time. MOA is well suited for these requirements. Weka and MOA can be closely linked to each other and either of the classifiers can be called from the other one. For those looking to analyse and mine information from real-time data, MOA can be the best choice.

KEEL: (Knowledge Extraction for Evolutionary Learning) is a Java based open source tool distributed under GPLv3. It is powered by a well-organised GUI that lets you manage (import, export, edit and visualise) data with different file formats, and to experiment with the data (through its data pre-processing, statistical libraries and some standard data mining and evolutionary learning algorithms). Since KEEL is based on Java, JVM has to be installed on the system to run its GUI and do data mining experiments. KEEL is ideal for research and educational purposes. It serves as a useful aid for teachers.

Rattle: Expanded to 'R Analytical Tool To Learn Easily', has been developed using the R statistical programming language. The software can run on Linux, Mac OS and Windows, and features statistics, clustering, modelling and visualisation with the computing power of R. Rattle is currently being used in business, commercial enterprises and for teaching purposes in Australian and American universities.

All the tools and software discussed so far are not the only available ones—the list keeps growing. While I have covered only those tools exclusively meant for mining data, there are a few other machine learning, NLP and data analytic tools that could aid in mining, like scikit-learn, NLTK, Graph Lab, Neural Designer, Pandas and SPMF, which readers could explore.

III.HISTORY OF KNIME

Konstanz Information Miner, is an open source data analytics, reporting and integration platform. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept. A graphical user interface allows assembly of nodes for data pre-processing (ETL: Extraction, Transformation, Loading), for the Development of KNIME was started January 2004 by a team of software engineers at University of Konstanz as a proprietary product. The original developer team headed by Michael Berthold came from a company in Silicon Valley providing software for the pharmaceutical industry. The initial goal was to create a modular, highly scalable and open data processing platform which allowed for the easy integration of different data loading, processing, transformation, analysis and visual exploration modules without the focus on any particular application area. The platform was

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

intended to be a collaboration and research platform and should also serve as an integration platform for various other data analysis projects.

INTERNALS: KNIME allows users to visually create data flows (or pipelines), selectively execute some or all analysis steps, and later inspect the results, models, and interactive views. KNIME is written in Java and based on Eclipse and makes use of its extension mechanism to add plugins providing additional functionality. The core version already includes hundreds of modules for data integration. (file I/O, database nodes supporting all common database management systems through JDBC), data transformation (filter, converter, combiner) as well as the commonly used methods for data analysis and visualization. With the free Report Designer extension, KNIME workflows can be used as data sets to create report templates that can be exported to document formats like doc, ppt, xls, pdf and others. Other capabilities of KNIME are:KNIMEs core-architecture allows processing of large data volumes that are only limited by the available hard diskspace (most other open source data analysis tools work in main memory and are therefore limited to the available RAM). E.g. KNIME allows analysis of 300 million customer addresses, 20 million cell images and 10 million molecular structures. Additional plugins allows the integration of methods for Text mining, Image mining, as well as time series analysis. KNIME integrates various other opensource projects, e.g. machine learning algorithms from Weka, the statistics package R project, as well as LIBSVM, JFreeChart, ImageJ, and the Chemistry Development Kit.KNIME is implemented in Java but also allows for wrappers calling other code in addition to providing nodes that allow to run Java, Python, Perl and other code fragments.

LICENSE: As of version 2.1, KNIME is released under GPLv3 with an exception that allows others to use the well-defined node API to add proprietary extensions. This allows also commercial SW vendors to add wrappers calling their tools from KNIME.

RATTLE: Rattle is a popular GUI for data mining using R. It presents statistical and visual summaries of data, transforms data so that it can be readily modelled, builds both unsupervised and supervised machine learning models from the data, presents the performance of models graphically, and scores new datasets for deployment into production. A key features is that all of your interactions through the graphical user interface are captured as an R script that can be readily executed in R independently of the Rattle interface. Use it as a tool to learn and develop your skills in R and then to build your initial models in Rattle to then be tuned in R which provides considerably more powerful options.

USERS:

Rattle is in daily use by Australia's largest team of data scientists and by a variety of government and other enterprises, worldwide. Whilst the true number of active users is hard to gauge we can observe that there are about 20,000 downloads of the package per month from a single popular CRAN node (where CRAN has over 100 nodes).

Many independent consultants worldwide also use Rattle in their day-to-day business.

Through various contacts we know that users include ANZ Bank, Commonwealth Bank, the Australian Taxation Office, Australian Department of Immigration, Ulster Bank, Toyota Australia, Vitorian Cancer Council, US Geological Survey, Carat Media Network, Institute of Infection and Immunity of the University Hospital of Wales, US National Institutes of Health, AIMIA Loyalty Marketing, Added Value, University of Canberra, Harbin Institute of Technology, Shenzhen Graduate School (since 2006), Australian Consortium for Social and Political Research (2011), Revolution Analytics (since 2012 and now Microsoft), International Centre for Free and Open Source Software in Kerala, India (2015) and many others.

INSTALL: From within R you can install Rattle directly from CRAN with:

R: > install. Packages ("rattle")

An alternative is to install the most recent release from Togaware:

R: > install. Packages ("rattle",

repos="http://rattle.togaware.com")

IV.CONCLUSION

Many advanced tools for data mining are available either as open-source or commercial software. They cover a wide range of software products, from comfortable problem-independent data mining suites, to business-centered data warehouses with integrated data mining capabilities, to early research prototypes for newly developed methods.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Website: www.ijirset.com

Vol. 6, Issue 8, August 2017

REFERENCES

1. Hahsler, M., Gr¨un, B. and Hornik, K. (2005),A Computational Environment for Mining Association Rules and Frequent Item Sets, R Package, Version 0.2-1.
2. **Graham Williams (2011). Data Mining with Rattle and R: The Art of Excavating Data for Knowledge Discovery, Springer.**
3. Rattle is described as an "attractive, easy-to-use front end ... data mining toolkit" in an article published in the Teradata Magazine, volume 9, issue 3, page 57 (September 2009).
4. Data Mining and Analysis: Fundamental concepts and Algorithms by Mohammed J. Zaki and Wagner Meira, Jr. Cambridge University Press, May 2014