# A Survey on Tools and Techniques in Data Mining

**Santhosh Pawar**
Department of CS
KSWU Vijayapura

***Abstract-****Data mining is the process of extracting the useful data, patterns and trends from a large amount of data by using techniques like clustering, classification, association and regression. Data mining techniques are used for information retrieval, statistical modelling and Machine learning. These techniques employ data pre-processing, data analysis and data interpretation. There are a wide variety of applications in real life. Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. Various tools are available which supports different algorithms. A summary about data mining tools available and the supporting algorithms is the objective of this paper. Comparison between various tools has also been done to enable the users use various tools according to their requirements and applications.*

***Keywords***-Data mining, Clustering, Classification, Association and Regression .

## I. INTRODUCTION

Data mining is the withdrawal of hidden predictive information from large databases. It allows us to find the needles hidden in our haystacks of data. It is a prevailing new technology which has great potential to help companies that focus on the most important information in their data warehouses. Tools of data mining predict future trends and behaviours, allowing businesses to make proactive and the knowledge-driven decisions. Data mining offers the automated, prospective analyses beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally were too time consuming to resolve. They search databases for hidden patterns, finding analytical information that experts may miss because it lies outside their expectations. It is playing increasingly important role in both private and public sectors. E.g. the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment. Data mining can be used in a predictive manner for a variety of applications.

Data mining need: Nowadays, large quantities of data are being accumulated. The amount of data collected is said to be almost doubled every 12 months. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. Data could be large in two senses. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. Data could be large in two senses. In terms of size, e.g. for Image Data or in terms of dimensionality, e.g. for Gene expression data. Usually there is a huge gap from the stored data to the knowledge that could be construed from the data. This transition won't occur automatically, that's where Data Mining comes into picture. In Exploratory Data Analysis, some initial knowledge is known about the data, but Data Mining could help in a more in-depth knowledge about the data. Manual data analysis has been around for some time now, but it creates a bottleneck for large data analysis. Fast developing computer science and engineering techniques and methodology generates new demands. Data Mining techniques are now being applied to all kinds of domains, which are rich in data, e.g. Image Mining and Gene data analysis.

Data Mining Techniques Clustering - Clustering is a process of partitioning a set of data (or objects) into a set of meaningful sub-classes, called clusters. Its main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression and computer graphics.

Classification – Classification is used to classify each item in a set of data into one of predefined set of classes or groups. The goal of classification is to accurately predict the target class for each case in the data.

Example: A bank loan officer wants to analyze the data in order to know which customers (loan applicant) are risky or which are safe.

## II. CLASSIFICATION ISSUES

Data Cleaning - Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.

Relevance Analysis - Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.

Data Transformation and reduction - The data can be transformed by any of the following methods.

i. Normalization − The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.

ii. Generalization − The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

Association Rule: An association rule has two parts, an antecedent (if) and a consequent (then). An antecedent is an item found in the data. A consequent is an item that is found in combination with the antecedent. Association rules are created by analyzing data for frequent if / then patterns and using the criteria support and confidence to identify the most important relationships. Regression: Regression is a data mining function that predicts a number. Age, weight, distance, temperature, income, or sales could all be predicted using regression techniques. For example, a regression model could be used to predict children's height, given their age, weight, and other factors. Data Mining Tools Data science combines data mining, machine learning and statistical methodologies to extract knowledge and leverage predictions from data. Given the need for data science in organizations, many small or medium organizations are not adequately funded to the current state of the art is necessary. This work explores and compares common open source data science tools. Implications include an overview of the state of the art and knowledge for practitioners and academics to select an open source data science tool that suits the requirements of specific data science projects. Open Source Tools for the Data Scientist This section covers current reviews on open source data science tools. Following the review, open source tools are compared based on the industry data science certification. Orange Tool Orange is an open source data mining, visualization environment, analytics, and scripting environment. Widgets are used as the building blocks to create workflows within the Orange environment. Widgets are categorized as Data, Visualize, Classify, Regression, Evaluate, Associate and Unsupervised. Tanagra Tool Tanagra claims to be an open source environment for teaching and research and is the successor to the SPINA software. Capabilities include Data source, Visualization, Descriptive statistics, Instance selection, Feature selection, Feature construction, Regression, Factorial analysis, Clustering, Supervised learning, Meta - Supervised learning, Learning assessment, and Association Rules. Rapid Miner Rapid Miner, formerly Yale, has morphed into a licensed software product as opposed to open source. Nevertheless, Rapid Miner community edition is still free and open source. Rapid Miner has the ability to perform process control (i.e. loops), connect to a repository, import and export data, data transformation, modelling (i.e. classification and regression), and Evaluation. KNIME KNIME is the Konstant Information Miner which had its beginnings at the University of Konstanz and has since developed into a full- scale data science tool. There are multiple versions of KNIME each with added capabilities. Much like Rapid Miner, advanced capabilities and tools come at a price. Functionalities include univariate and multivariate statistics, data mining, time series analysis, image processing, web analytics, text mining, network analysis, and social media analysis.

R R is a free and open source package for statistics and graphing. R is traditionally command line; however, there are many feely available open source tools that integrate into R. One such example is R Studio which provides a graphical user interface for R. R can be employed for a variety of statistical and analytics tasks including but not limited to clustering, regression, time series analysis, text mining, and statistical modelling. R is considered an interpreted language more so than an environment. R supports big data processing with RHadoop. RHadoop connects R to Hadoop environments and runs R programs across Hadoop nodes and clusters. Weka Weka, or the Waikato Environment for Knowledge Analysis, is licensed under the GNU general public license. Weka stems from the University of Waikato and is a collection of packages for machine learning and is Java based. Weka provides an API so developers may use Weka from their projects. Weka is widely adopted in academic and business and has an active community. Weka's community has contributed many add-in packages such as K – anonymity and I – diversity for privacy preserving data mining and bagging and boosting of decision trees. Tools may be downloaded from a repository and via the package manager. Weka is java based and extensible. Weka provides .jar files which may be built into any Java application permitting custom programming outside of the Weka environment. For big data processing, Weka has its own packages for map reduce programming to maintain independence over platform but also provides wrappers for Hadoop.

## III. CONCLUSION

In this paper, we have discussed detail study of data mining with various studies like tasks, tools and techniques. The implementation of data mining techniques will allow

users to retrieve meaningful information from virtually integrated data. These techniques provide variety of applications for industries like retail, telecommunication, bio-medical etc. This research also conducted a comparison between four data mining toolkits for classification purposes. However, in terms of classifiers applicability, we concluded that the WEKA toolkit was the best tool in terms of the ability to run the selected classifier followed by Orange, Tanagra, and finally KNIME respectively.

## REFERENCES

[1] Srivastava S (2014) Weka: A Tool for Data preprocessing, Classification, Ensemble, Clustering and Association Rule Mining. International Journal of Computer Applications (0975 –8887).

[2] Aviad B, Roy G (2012) A decision support method, based on bounded rationality concepts, to reveal feature saliency in clustering problems. Decision Support Systems 54: 292–303 .

[3] Combes C, Azema J (2013) Clustering using principal component analysis applied to Autonomy disability of elderly people. Decision Support Systems 55:578–586

[4] Rao GN, Ramachandra M (2014) Study on the Academic Performance of the Students by Applying K-Means Algorithm. IJETCAS 14-180.

[5] Sheetal L. Patil "Survey of Data Mining Techniques in Health-care" International Research Journal of Innovative Engineering Volume1, Issue 9 of September 2015.

[6] Demšar J, Zupan B (2013) Orange: Data Mining Fruitful and Fun -A Historical Perspective. Informatica 37:55–60.

[7] Radaideh Q, Nagi E (2012) Using Data Mining Techniques to Build a Classification Model for Predicting Employees Performance. IJACSA

[8] Velmurugan T (2014) Performance based analysis between k-Means and Fuzzy CMeans clustering algorithms for connectionoriented telecommunication data. Applied Soft Computing 19 pp.134–146.

[9] Dr.Varun Kumar, Anupama Chadha, "Mining Association Rules in Student's Assessment Data", IJCSI International Journal of Computer Science Issues,Vol.Issue 5, No 3, September 2012.

[10] Divya, S. A. (2011, June 3-4). Weighted Support Vector Regression approach for Remote Healthcare monitoring. IEEE-International Conference on Recent Trends in Information Technology(pp. 978-1-4577). Chennai: 0590-8/11/$26.00 © 2011 IEEE MIT