A Study of Data Mining Techniques and Tools in the Process of Knowledge Discovery

[1] Prerna S. Tayade, [2] Dr. S.R. Kalmegh,
[1]Research Scholar, [2]Associate Professor
Computer Science Department
[1]Shri Shivaji Science College, Amravati, India, [2]SGBAU Amravati, India.

Abstract: In the real world we are dealing with the large amount of data and we want to extract the meaningful information among that data. Data mining is also known as Knowledge discovery and data mining. Rapid computerization of business produces huge amount of data so how to make best use of that data and how to derive patterns from this data mining is important. The extracted patterns must be valid, novel, useful and understandable. Thus, data mining provides the way for extraction of hidden predictive information from huge databases. It is the process of finding previously unknown and potentially useful information from databases. Data mining techniques have highly useful in medical field as there is voluminous data in this industry. Data mining uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. Number of data mining tools are available today. This paper presents an overview of the data mining tools like KMIME, WEKA, ORANGE, RAPID MINER. Also focus is given on the summary of data mining techniques used for all the domains.

***Index Terms* - Data mining, Knowledge Discovery Database.**

I. INTRODUCTION

The data sources such as data warehouse, database, flat files and other data repositories contain a huge amount of data and information from which it is impossible to identify an useful one for good decision making process. Data mining is the process of discovering interesting knowledge from large amounts of data stored in databases, data warehouses, or other information repositories. Such kind of knowledge discovery led to the development of data mining tools which will assist us in transforming those vast amounts of data into useful information and knowledge.[8] Various data mining tools are now available in market. one has to find that which tool is best suits for his organization for a successful decision-making process. This paper reviews data mining Techniques and various free open source tools available for commercial purpose.

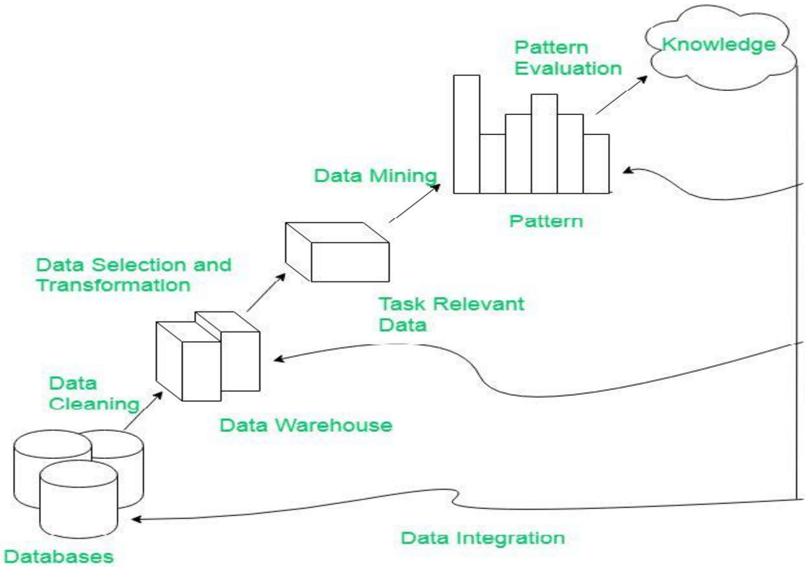

Figure 1: Data mining is the core of Knowledge discovery process.

II. Techniques in Data Mining:

a) Classification:

Classification is one of the most important technique in data mining. It is based on machine learning. Generally classification is use to classify the data into groups. Classification allows us to collect various attributes together into different categories. which we can use to perform different functions and may use in future decision making.[6] For example, if you are interested in individual customer's financial background and purchase history, you might be able to classify them as "low," "medium," or "high" credit risks. Then you can use these classifications to know more about these customers. Classification technique use some methods for classification such as decision trees, linear programming, neural network, and statistics.

b) Association:

Association is an important data mining technique which is related to tracking patterns. The patterns are discovered based on relationship between items in same transaction. The association technique is generally used in market analysis. It includes the study of customer behavior, that purchase habit of customer as which products customer buy together. As Association is use in tracking pattern it is dependent on related variables. [6]For e.g. Generally, it may notice that if your customer buy some product they may also interested to buy certain related product. This is usually showing in the section of people also bought in online shopping sites. This is based on the purchase behavior of customer.

c) Clustering:

Clustering is a data mining technique it is similar to classification technique as it includes grouping of data, but the grouping is done on the basis of similarity. That means in clustering clusters are form on the basis of similar characteristics of objects. In clustering technique classes are defined and objects are assigned to each class depends on similarity.[8] For eg. As we take a scenario of library, in the library there is a huge collection of books of different subjects. And if we want to search a book of particular subject it is very difficult to search the book in unstructured form. Using clustering this task get easy as if we group the book of related similarity and place it in the separate shelf and name the shelf to specific subject then it is easy to us to find the particular book related to particular subject. . Clustering mainly used for pattern recognition, machine learning and information retrieval.

d) Decision trees

A decision tree is a flow-chart-like tree structure, where each node denotes a test on an attribute value, each branch represents an outcome of the test, and tree leaves represent classes or class distributions. It is a predictive model that maps observations about an item to conclusions about the item's target value.[6] They can be easily converted to classification rules. It can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions.

e) Outlier detection

In many cases, simply recognizing the overarching pattern can't give you a clear understanding of your data set. You also need to be able to identify anomalies, or outliers in your data.[6] For example, if your purchasers are almost exclusively male, but during one strange week in July, there's a huge spike in female purchasers, you'll want to investigate the spike and see what drove it, so you can either replicate it or better understand your audience in the process.

f) Prediction

Prediction is one of the most valuable data mining techniques, since it's used to project the types of data you'll see in the future. In many cases, just recognizing and understanding historical trends is enough to chart a somewhat accurate prediction of what will happen in the future. For example, you might review consumers' credit histories and past purchases to predict whether they'll be a credit risk in the future[8].

g) 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.[8]

III. Tools in Data Mining

a) KNIME:

Konstanz Information Miner (KNIME) is one of the open source data mining tool. This acme software was developed at the University of Konstanz headed by Michael Berthold from Silicon Valley in January 2014.

KNIME (Konstanz Information Miner) is a user friendly, knowledgeable and comprehensive data integration, processing, analysis, and exploration platform. It provides the users to create data flows or pipelines visually, users can selectively execute some or all analysis steps, study the results, prototypes, and collaborative interpretations. KNIME is written in Java, and based on Eclipse.[1]

b) Orange:

Orange is a powerful free and open source component-based data mining and machine learning software suite. It contains complete set of components for data preprocessing, feature scoring and filtering, modeling, model evaluation, and exploration techniques. It is based on C++ components, that are accessed either directly (not very common), through Python scripts (easier and better), or through GUI objects called Orange Widgets. Orange is distributed free under GPL and can be downloaded from the download page. Orange is a component-based framework, which means you can use existing components and build your own ones. You can even prototype your own components in Python, and use it in place of some standard C-based Orange component. Orange is supported on various versions of Linux, Apple's, Mac OS X and Microsoft Windows.[7]

c) RapidMiner

RapidMiner is a software platform developed by the rapid miner company. It is formerly known as YALE (Yet Another Learning Environment). It uses a client/server model with the server as the Software as a Service (SAAS) or on a cloud infrastructure. It provides a unified background for machine learning, data mining, text mining, predictive analytics and business analytics. It is mainly used for business and industrial applications as well as for research, education, training, rapid prototyping, and application development and supports all the steps of the KDD including results such as visualization, validation and optimization.[7]

d) WEKA

Waikato Environment for Knowledge Analysis is developed at the University of Waikato (Hall et al., 2009), New Zealand. It provides toolkits for machine learning. WEKA provides a comprehensive collection of algorithms. WEKA provides itself with the GUI, so algorithms can be easily used to the dataset from GUI directly or it can be called from the Java code. It also supports working in the command prompt. Tasks like preprocessing, feature selection, clustering, can be done using WEKA. It is written in Java so it is platform independent and can run in almost any platform. It also supports visualization tasks and many machine learning applications. WEKA is freeware available under the General public license agreement (GNU).[1]

Table 8: Classification algorithms supported by KNIME, WEKA, ORANGE and Rapid Miner[2]

	Rapid Miner	WEKA	KNIME	ORANGE
K-means	✓	✓	✓	✓
x-means	✓	✓		
DBSCAN	✓			
Expectation maximization cluster	✓			
Support vector clustering	✓			
Random clustering	✓			
Agglomerative clustering	✓			
Top down clustering	✓			
Flatten clustering	✓			
Extract cluster prototypes	✓			
Cobweb		✓		
Farthest first		✓		
Filtered clusterer		✓		
Hierarchical clusterer		✓		✓
OPTICS		✓		
Sib		✓		
Density-based clustering algorithm		✓		

IV. Conclusion

This study has given a brief introduction on the five different data mining tools and their applications KNIME, ORANGE, WEKA and Rapid Miner. Each tool has its own pros and cons. Be that as it may, KNIME, ORANGE, WEKA and Rapid Miner have most of the desired characteristics and functions for a fully-functional Data Mining platform and thereby these tools can be used for most of the Data Mining tasks. As a future work, we are going to study and compare the performance of various data mining classification and clustering algorithms for various data mining tools.

References

1] S.Singh, Y.Liu, W.Ding and Z.Li,"Evaluation of data mining tools for Telecommunication Monitoring Data using design of experiment," IEEE ,2016.
2] M.Hassan , ME.Shahab , EMR.Hamed.,"A comparative study of classification algorithm in E-health Environment," IEEE.2016.
3] Top 10 challenging problems in Data mining[Online].Available from: http://www.dataminingblog.com/top-10-challengingproblems-in-data-mining/
4] H.Odan, A.Daraiseh,"Open source Data Mining Tools," IEEE,2015.
5] R.Arun and J.Tamilselvi,"Data Quality and the Performance of the Data Mining Tool",2015.
6] G.Keseavaraj, S.Sukumaran,"Study on classification techniques on data mining," 4th ICCCNT ,IEEE, 2013.
7] R.Mikut and M.Reischl,"Data mining tools. Research gate,"2011
8] H. Jiawei , M. Kamber, J. Pei, Data mining concepts and techniques, 3rd ed., Morgan Kaufmann Elsevier: USA , 2012