10th International Conference on Recent Development in Engineering Science, Humanities and Management
The Indian Council of Social Science Research (ICSSR) North West Regional Centre , Punjab University Campus, Chandigarh (India) RDESHM-17
24th December 2017, www.conferenceworld.in ISBN: 978-93-86171-89-4

# Study and Applications of Data Mining Tools and Techniques

**Meenakshi**

**ABSTRACT**

*Banking frameworks gather enormous measures of information on everyday premise, be it client data, exchange points of interest, chance profiles, Mastercard subtle elements, utmost and guarantee points of interest, consistence and Anti Money Laundering (AML) related data, exchange back information, SWIFT and wire messages. Associations have been currently executing information warehousing innovation, which encourages colossal venture wide databases. Subsequently, the measure of information that associations have is developing at an incredible rate. The following test for these associations is the means by which to decipher the information and how to change it into valuable data and learning. Data mining is one innovation utilized for addressing this difficulty. In this article, the author discuss about the Applications of Data Mining Tools and Techniques*.

**Keywords: Data mining, software, applications, tools.**

## I.INTRODUCTION

Data mining is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both. Data mining software is one of a number of analytical tools for analyzing data. It enables clients to dissect information from a wide range of measurements or edges, sort it, and condense the connections distinguished. In fact, information mining is the way toward discovering connections or examples among many fields in extensive social databases. Information mining is essentially utilized today by organizations with a solid shopper center - retail, monetary, correspondence, and promoting associations. It empowers these organizations to decide connections among "inward" factors, for example, value, item situating, or staff aptitudes, and "outer" factors, for example, financial markers, rivalry, and client socioeconomics. Furthermore, it empowers them to decide the effect on deals, consumer loyalty, and corporate benefits. At last, it empowers them to "penetrate down" into synopsis data to see detail value-based information.

Information mining alludes to removing learning from a lot of information. The information might be spatial information, sight and sound information, time arrangement information, content information and web information. Information mining is the procedure of extraction of fascinating, nontrivial, certain, already obscure and conceivably helpful examples or learning from enormous measures of information. It is the arrangement of exercises used to discover new, covered up or startling examples in information or strange examples in information. Utilizing data contained inside information distribution center, information mining can

regularly give answers to inquiries concerning an association that a chief has already not thought to ask . With information mining, a retailer could utilize purpose of-offer records of client buys to send focused on advancements in view of a person's buy history. By mining statistic information from remark or guarantee cards, the retailer could create items and advancements to engage particular client portions.

## II. DATA MINING TOOLS

Data mining is not all about the tools or database software that you are using. You can perform information mining with similarly humble database frameworks and straightforward apparatuses, including making and composing your own, or utilizing off the rack programming bundles. Complex information mining profits by the past experience and calculations characterized with existing programming and bundles, with specific devices picking up a more prominent fondness or notoriety with various methods. For instance, IBM SPSS®, which has its underlying foundations in measurable and overview investigation, can fabricate compelling prescient models by taking a gander at past patterns and building exact conjectures. IBM InfoSphere® Warehouse gives information sourcing, preprocessing, mining, and examination data in a solitary bundle, which enables you to take data from the source database straight to the last report yield.

It is later that the extensive informational indexes and the bunch and expansive scale information handling can permit information mining to order and write about gatherings and connections of information that are more entangled. Presently a totally new scope of apparatuses and frameworks accessible including joined information stockpiling and handling frameworks. You can mine information with a different distinctive informational collections, including, customary SQL databases, crude content information, key/esteem stores, and report databases. Grouped databases, for example, Hadoop, Cassandra, CouchDB, and Couchbase Server, store and give access to information so as to not coordinate the conventional table structure. Specifically, the more adaptable stockpiling configuration of the record database causes an alternate concentration and many-sided quality as far as preparing the data. SQL databases impost strict structures and unbending nature into the pattern, which makes questioning them and breaking down the information direct from the point of view that the configuration and structure of the data is known. Report databases that have a standard, for example, JSON authorizing structure, or documents that have some machine-lucid structure are likewise simpler to process, in spite of the fact that they may include complexities in view of the varying and variable structure. For instance, with Hadoop's completely crude information preparing it can be perplexing to distinguish and extricate the substance before you begin to process and relate it.

It is legitimately said that information is cash in this day and age. Alongside the change to an application based world comes the exponential development of information. In any case, a large portion of the information is unstructured and henceforth it takes a procedure and technique to separate helpful data from the information and change it into justifiable and usable frame. This is the place information mining comes into picture. A lot of apparatuses are accessible for information mining assignments utilizing counterfeit consciousness, machine

learning and different methods to remove information. Here are six effective open source information mining devices accessible:

**a. Rapidminer:**

Written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. A bonus: Users hardly have to write any code. Offered as a service, rather than a piece of local software, this tool holds top position on the list of data mining tools. In addition to data mining, Rapid Miner also provides functionality like data preprocessing and visualization, predictive analytics and statistical modeling, evaluation, and deployment. What makes it even more powerful is that it provides learning schemes, models and algorithms from WEKA and R scripts. RapidMiner is distributed under the AGPL open source license and can be downloaded from Source Forge where it is rated the number one business analytics software.

**b. WEKA:**

The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modeling. Its free under the GNU General Public License, which is a big plus compared to RapidMiner, because users can customize it however they please.

WEKA supports several standard data mining tasks, including data preprocessing, clustering, classification, regression, visualization and feature selection. WEKA would be more powerful with the addition of sequence modeling, which currently is not included.

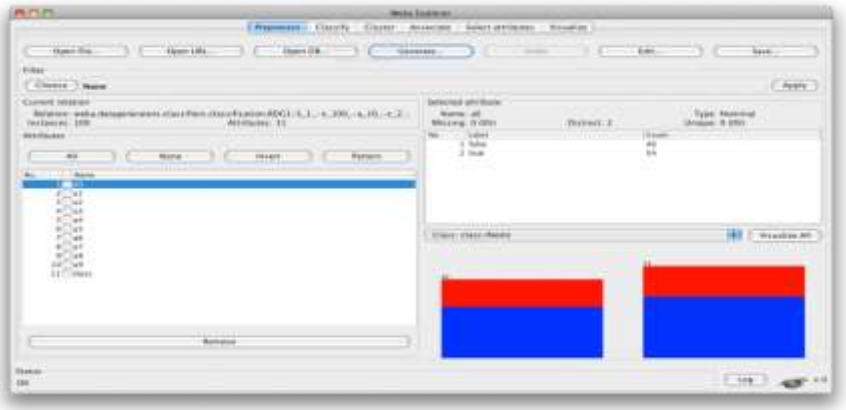

**Figure 1: WEKA Browser**

**c. R-Programming:**

What if I tell you that Project R, a GNU project, is written in R itself? It's primarily written in C and Fortran. And a lot of its modules are written in R itself. It's a free software programming language and software environment for statistical computing and graphics. The R language is widely used among data miners for developing statistical software and data analysis. Ease of use and extensibility has raised R's popularity substantially in recent years. Besides data mining it provides statistical and graphical techniques, including linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others.

**d. Orange:**

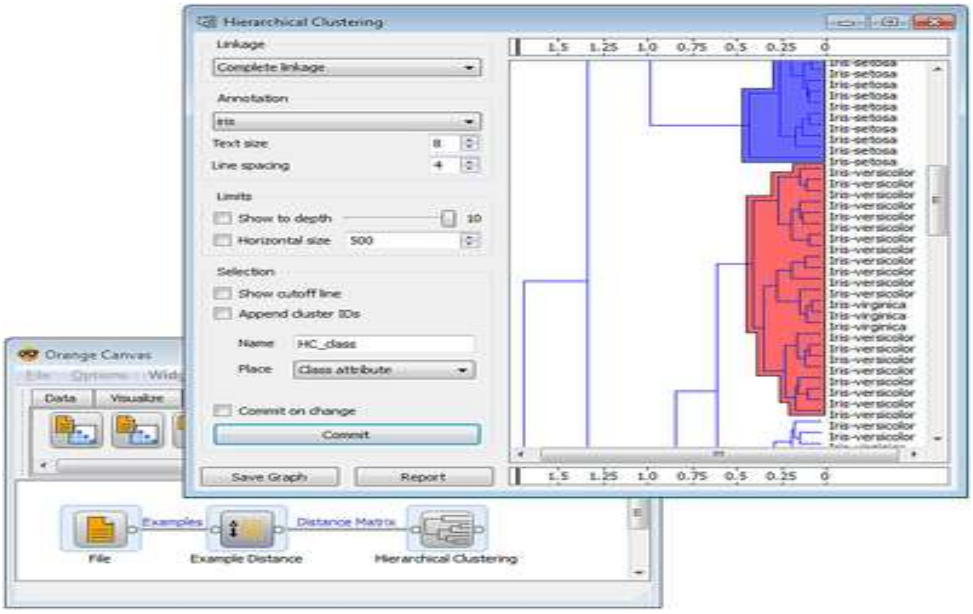


**Fig. 2: Orange Hierarchical Clustering**

Python is picking up in popularity because it's simple and easy to learn yet powerful. Hence, when it comes to looking for a tool for your work and you are a Python developer, look no further than Orange, a Python-based, powerful and open source tool for both novices and experts. You will fall in love with this tool's visual programming and Python scripting. It also has components for machine learning, add-ons for bioinformatics and text mining. It's packed with features for data analytics.

**e. KNIME:**

Data preprocessing has three main components: extraction, transformation and loading. KNIME does all three. It gives you a graphical user interface to allow for the assembly of nodes for data processing. It is an open source data analytics, reporting and integration platform. KNIME also integrates various components for machine learning and data mining through its modular data pipelining concept and has caught the eye of business intelligence and financial data analysis. Written in Java and based on Eclipse, KNIME is easy to extend and to add plugins. Additional functionalities can be added on the go. Plenty of data integration modules are already included in the core version.

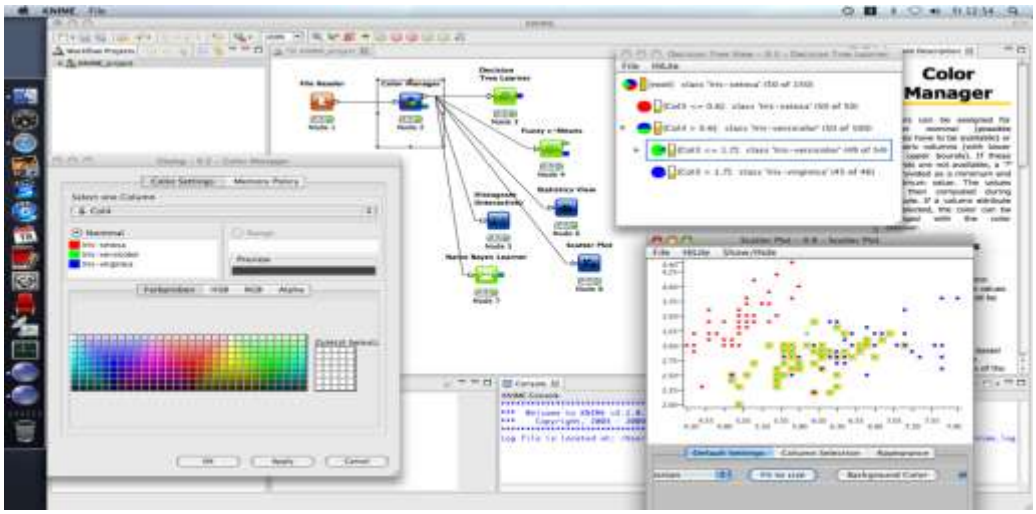


**Fig. 3: KNIME Browser**

**f. NLTK:**

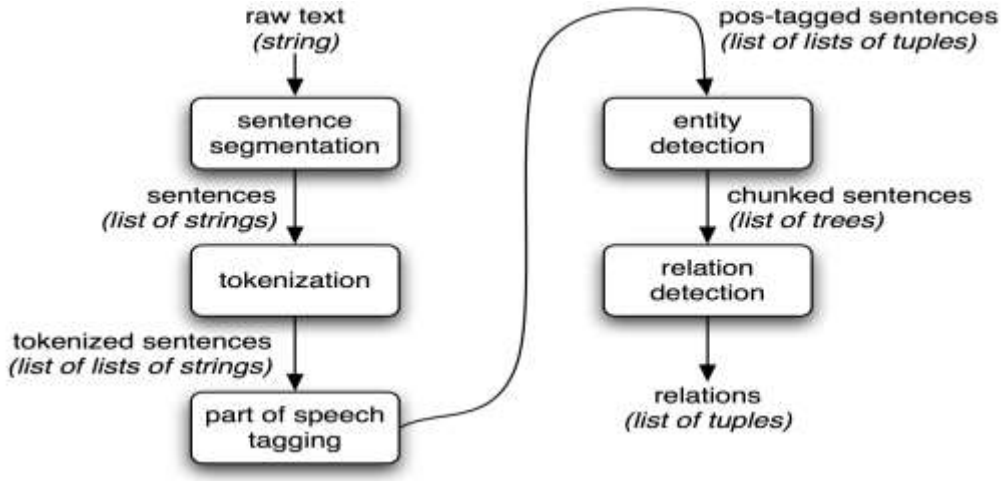


**Fig. 4: NLTK Tool**

When it comes to language processing tasks, nothing can beat NLTK. NLTK provides a pool of language processing tools including data mining, machine learning, data scraping, sentiment analysis and other various language processing tasks. All you need to do is install NLTK, pull a package for your favorite task and you are ready to go. Because it's written in Python, you can build applications on top if it, customizing it for small tasks.

## III. DATA MINING TECHNIQUES

There are several major data mining techniques have been developing and using in data mining projects recently including association, classification, clustering, prediction, sequential patterns and decision tree. We will quickly look at those information mining strategies in the accompanying segments.

**Association**

Association is a standout amongst other known information mining strategies. In affiliation, an example is found in light of a connection between things in a similar exchange. That is the motivation behind why affiliation system is otherwise called connection procedure. The affiliation strategy is utilized as a part of market container examination to recognize an arrangement of items that clients often buy together. Retailers are utilizing affiliation procedure to inquire about client's purchasing propensities. In light of chronicled deal information, retailers may discover that clients dependably purchase crisps when they purchase brews, and, in this manner, they can put lagers and crisps alongside each other to spare time for client and increment deals.

**Classification**

Classification is a great information mining procedure in view of machine learning. Essentially, order is utilized to characterize everything in an arrangement of information into one of a predefined set of classes or gatherings. Grouping strategy makes utilization of numerical strategies, for example, choice trees, direct programming, neural system and insights. In characterization, we build up the product that can figure out how to order the information things into gatherings. For instance, we can apply arrangement in the application that "given all records of workers who left the organization, foresee who will most likely leave the organization in a future period." For this situation, we partition the records of representatives into two gatherings that named "leave" and "remain". And after that we can ask our information mining programming to arrange the representatives into partitioned gatherings.

**Clustering**

Clustering is an information mining system that makes a significant or helpful bunch of articles which have comparable qualities utilizing the programmed strategy. The bunching system characterizes the classes and

places questions in each class, while in the grouping strategies, objects are appointed into predefined classes. To make the idea clearer, we can take book administration in the library for instance. In a library, there is an extensive variety of books on different subjects accessible. The test is the manner by which to keep those books in a way that perusers can take a few books on a specific theme without issue. By utilizing the bunching system, we can keep books that have a few sorts of likenesses in a single group or one retire and name it with an important name. On the off chance that perusers need to snatch books in that subject, they would just need to go to that rack as opposed to searching for the whole library.

**Prediction**

The Prediction, as its name suggested, is one of an information mining systems that find the connection between free factors and connection amongst needy and autonomous factors. For example, the expectation investigation method can be utilized as a part of the deal to anticipate benefit for the future on the off chance that we consider the deal is an autonomous variable, benefit could be a reliant variable. At that point in light of the chronicled deal and benefit information, we can draw a fitted relapse bend that is utilized revenue driven forecast.

**Successive Patterns**

Consecutive examples examination is one of information mining method that tries to find or recognize comparable examples, normal occasions or patterns in exchange information over a business period.In deals, with chronicled exchange information, organizations can distinguish an arrangement of things that clients purchase together extraordinary circumstances in a year. At that point organizations can utilize this data to prescribe clients get it with better arrangements in light of their obtaining recurrence before.

**Decision trees**

A decision tree is a standout amongst the most well-known utilized information mining systems since its model is straightforward for clients. In choice tree procedure, the foundation of the choice tree is a basic inquiry or condition that has different answers. Each answer at that point prompts an arrangement of inquiries or conditions that assistance us decide the information with the goal that we can settle on an official conclusion in light of it. For instance, We utilize the accompanying choice tree to decide if to play tennis:

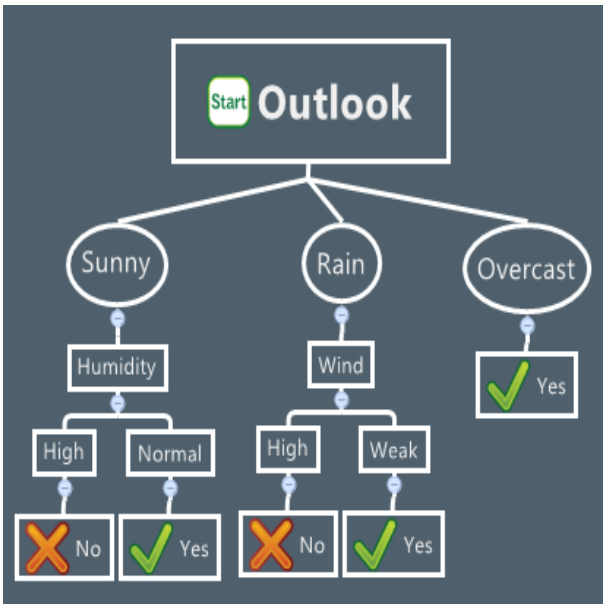


**Fig. 4: NLTK Tool**

Starting at the root node, if the outlook is overcast then we should definitely play tennis. If it is rainy, we should only play tennis if the wind is the week. And if it is sunny then we should play tennis in case the humidity is normal. We often combine two or more of those data mining techniques together to form an appropriate process that meets the business needs.

## IV.CONCLUSION

Data Mining techniques can be of immense help to the banks and financial institutions in this arena for better targeting and acquiring new customers, fraud detection in real time, providing segment based products for better targeting the customers, analysis of the customers' purchase patterns over time for better retention and relationship, detection of emerging trends to take proactive approach in a highly competitive market adding a lot more value to existing products and services and launching of new product and service bundles. Information mining is more than running some intricate questions on the information you put away in your database. You should work with your information, reformat it, or rebuild it, paying little heed to whether you are utilizing SQL, report based databases, for example, Hadoop, or basic level records. Recognizing the configuration of the data that you require depends on the procedure and the examination that you need to do. After you have the data in the arrangement you require, you can apply the distinctive procedures (exclusively or together) paying little heed to the required basic information structure or informational index.

**REFERENCES**

[1] Bhambri, V., 2011. Application of data mining in banking sector. IJCST, 2: 199-202.

[2] Bhattacharya, S., S. Jha, K. Tharakunnel and J.C. Westland, 2011. Data mining for credit card fraud: A comparative study. Decision Support Syst., 50: 602- 613. DOI: 10.1016/j.dss.2010.08.008.

[3] Babcock C. (1994) Parallel Processing Mines Retail Data. Computer World.

[4] Berry M.J.A. & Linoff G. (1999) Mastering Data Mining: The Art and Science of Customer Relationship Management. John Wiley and Sons, Inc.

[5] Davenport T.H. & Prusak L. (2000). Working Knowledge: How organizations manage what they know. Boston, Massachusetts; Harvard Business School Press.

[6] B. Desai and Anita Desai, "The Role of Data mining in Banking Sector", IBA Bulletin, 2004.

[7] S.S.Kaptan, "New Concepts in Banking", Sarup and Sons, Edition, 2002

[8] S. S. Kaptan, N S Chobey, "Indian Banking in Electronic Era", Sarup and Sons, Edition 2002.

[9] Rajanish Dass, "Data Mining in Banking and Finance: A Note for Bankers", Indian Institute of Management Ahmadabad.

[10] Moradi, M., M. Salehi, M.E. Ghorgani and H.S. Yazdi, 2013. Financial distress prediction of Iranian companies by using data mining techniques. Organizacija, 46: 20-27.

[11] Petry, F.E. and L. Zhao, 2009. Data mining by attribute generalization with fuzzy hierarchies in fuzzy databases. Fuzzy Sets Syst., 160: 2206-2223. DOI:10.1016/j.fss.2009.02.014.]

[12] Shinde, P., 2012. Data mining using artificial neural network tree. IOSR J. Eng. Tremblay, M.C., K. Dutta and D. Vandermeer, 2010.

[13] Using data mining techniques to discover bias patterns in missing data. J. Data Inform. Q., 2: 1-19.