Data Mining Tools And Applications

By Mrs. Priyanka Bhardwaj

Introduction to Data Mining

The world today runs on data but the question remains: How are we sourcing this data? The answer to it is not very simple. Data is often extracted from several sources and then crunched into useful numbers which are then used by companies to make any particular decision. Data Mining refers to the set of techniques and methods which involve the use of multiple software for the extraction and finding of suitable data points.

This involves a lot of research and brainstorming as suitable information has to be mined out of several sources. Data mining is not limited to finding out the sources of data. It majorly involves analyzing the existing databases for new insights as a particular sample set of data can give out a lot of new perspectives of looking at it. Data mining makes use of various statistics principles to trace the relationship between any two given data points. Also, in today's era, it uses tools like machine learning and artificial intelligence to find out various patterns and trends which become difficult to do manually.

What can be Achieved with Data Mining?

Data mining involves scrutinizing of raw data. This raw data can be anything like the price of a particular commodity at a particular point of time, the data on competitors, data on consumption pattern of a particular market segment, data on what marketing strategies have worked out well in a particular industry, etc. Data mining is also used for making business forecasts and predicting the uncertainties which envelop various business entities working in different sectors.

Also, the trend analysis is one of the major benefits of data mining as these give out patterns which help in analyzing the data faster and better and also gives powerful insights for a better understanding of the consumer base and their behavior such as the purchasing pattern, the kind of goods in which the consumer likes to spend, consumption patterns, frequency of purchases made by the customer, etc. thus leading to a productive and curated strategy to be taken to handle a particular market.

Also, with data mining, various hidden facts can be brought to the attention which will help the business in a lot of ways. For example: If a company is looking forward to entering a particular market segment, it would need a lot of information such as the size of the market, the size of the market which the company can tap, demand for that particular product in a particular area, etc. All this can be found out by hitting the eye and mining the correct and the most relevant data sources.

With the help of data mining, you avoid ambiguity and analyze data to extract information which is relevant to a particular business. With data mining, the operational costs can be brought done to a huge extent and with various automated tools available to mine data, the manpower cost has gone down drastically.

Data Mining Techniques

1. **Statistics:** Statistics deal with the collection and segmentation of data. Here, the quantitative aspect of data is being taken care of. This is an old technique that makes the trend analysis easy. Statistics bring various measures into the picture like regression, correlation, etc.

2. **Clustering:** Clustering of data is one of the most primitive and important steps in mining data. By this technique, the data is segregated into similar chunks and is divided into various segments which are then analyzed independently and also compared to the other segments thus formed.
3. **Visualization:** The visualization of data is a very important aspect of data mining. You can mine a lot of information from a given set of data but it is of no use when the person for whom the information is meant for is unable to understand it. It sanitizes the data and converts it into an understandable form that serves the purpose of data mining.
4. **Decision Tree:** Here, the data is arranged in the form of a tree showing the hierarchal and chronological relevance of different sets of data. Each branch of the tree is a classification and the data which supports the classification. This makes it easier for the user to make decisions and predictions.
5. **Association:** This technique aims at finding various links between two different sets of data or between various classifications made in the same data set. It establishes a relationship between various variables thus extracting valuable information for analysis and implementation.
6. **Neural Networks:** This is a basic foundation step which is automated. The user does not have to put in a lot of effort into the mining of data using neural networks. It is easy to use.
7. **Classification:** This is one of the most popular techniques used in mining data. Here, there are predefined classifications and models which classify a big set of data. It also brings in the element of other techniques which makes the data mining process a lot easier.

List Of Most Popular Data Mining Tools And Applications

Here we go!

Here we have compared the list of free and commercial data modeling tools.

#1) Xplenty

Xplenty provides a platform that has functionalities to integrate, process, and prepare data for analytics. Businesses will be able to make most of the opportunities offered by big data with the help of Xplenty and that too without investing in related personnel, hardware, and software. It is a complete toolkit for building data pipelines.

You will be able to implement complex data preparation functions through rich expression language. It has an intuitive interface to implement ETL, ELT, or a replication solution. You will be able to orchestrate and schedule pipelines through a workflow engine.

- Xplenty is the data integration platform for all. It offers the no-code and low-code options.
- An API component will provide advanced customization and flexibility.
- It has functionalities to transfer and transform data between databases and data warehouses.
- It provides support through email, chat, phone, and an online meeting.

Availability: Licensed tools.

#2) Rapid Miner

Availability: Open source
Rapid Miner is one of the best predictive analysis system developed by the company with the same name as the Rapid Miner. It is written in JAVA programming language. It provides an integrated environment for deep learning, text mining, machine learning & predictive analysis.

The tool can be used for over a vast range of applications including for business applications, commercial applications, training, education, research, application development, machine learning.

Rapid Miner offers the server as both on premise & in public/private cloud infrastructures. It has a client/server model as its base. Rapid Miner comes with template based frameworks that enable speedy delivery with reduced number of errors (which are quite commonly expected in manual code writing process).

Rapid Miner constitutes of three modules, namely

1. Rapid Miner Studio: This module is for workflow design, prototyping, validation etc.
2. Rapid Miner Server: To operate predictive data models created in studio
3. Rapid Miner Radoop: Executes processes directly in the Hadoop cluster to simplify predictive analysis.

#3) Orange

Availability: Open source
Orange is a perfect software suite for machine learning & data mining. It best aids the data visualization and is a component based software. It has been written in Python computing language.

As it is a component-based software, the components of orange are called ‘widgets'. These widgets range from data visualization & pre-processing to an evaluation of algorithms and predictive modeling.

Widgets offer major functionalities like

- Showing data table and allowing to select features
- Reading the data
- Training predictors and to compare learning algorithms
- Visualizing data elements etc.

Additionally, Orange brings a more interactive and fun vibe to the dull analytic tools. It is quite interesting to operate.

Data coming to Orange gets quickly formatted to the desired pattern and it can be easily moved where needed by simply moving/flipping the widgets. Users are quite fascinated by Orange. Orange allows users to make smarter decisions in short time by quickly comparing

---

#4) Weka

Availability: Free software
Also known as Waikato Environment is a machine learning software developed at the University of Waikato in New Zealand. It is best suited for data analysis and predictive modeling. It contains algorithms and visualization tools that support machine learning. Weka has a GUI that facilitates easy access to all its features. It is written in JAVA programming language.

Weka supports major data mining tasks including data mining, processing, visualization, regression etc. It works on the assumption that data is available in the form of a flat file.

Weka can provide access to SQL Databases through database connectivity and can further process the data/results returned by the query.

#5) KNIME

Availability: Open Source
KNIME is the best integration platform for data analytics and reporting developed by KNIME.com AG. It operates on the concept of the modular data pipeline. KNIME constitutes of various machine learning and data mining components embedded together.

KNIME has been used widely for pharmaceutical research. In addition, it performs excellently for customer data analysis, financial data analysis, and business intelligence.

KNIME has some brilliant features like quick deployment and scaling efficiency. Users get familiar with KNIME in quite lesser time and it has made predictive analysis accessible to even naive users. KNIME utilizes the assembly of nodes to pre-process the data for analytics and visualization.

#6) Sisense

Availability: Licensed
Sisense is extremely useful and best suited BI software when it comes to reporting purposes within the organization. It is developed by the company of same name ‘Sisense'. It has a brilliant capability to handle and process data for the small scale/large scale organizations.

It allows combining data from various sources to build a common repository and further, refines data to generate rich reports that get shared across departments for reporting.

Sisense got awarded as best BI software is 2016 and still, holds a good position.
Sisense generates reports which are highly visual. It is specially designed for users that are non-technical. It allows drag & drop facility as well as widgets.

Different widgets can be selected to generate the reports in form of pie charts, line charts, bar graphs etc. based on the purpose of an organization. Reports can be further drilled down by simply clicking to check details and comprehensive data.

#7) SSDT (SQL Server Data Tools)

Availability: Licensed
SSDT is a universal, declarative model that expands all the phases of database development in the Visual Studio IDE. BIDS was the former environment developed by Microsoft to do data analysis and provide business intelligence solutions. Developers use SSDT transact- a design capability of SQL, to build, maintain, debug and refactor databases.

A user can work directly with a database or can work directly with a connected database, thus, providing on or off-premise facility.

Users can use visual studio tools for development of databases like IntelliSense, code navigation tools, and programming support via C#, visual basic etc. SSDT provides **Table Designer** to create new tables as well as edit tables in direct databases as well as connected databases.
Deriving its base from BIDS, which was not compatible with Visual Studio2010, the SSDT BI came into existence and it replaced BIDS.

#8) Apache Mahout

Availability: Open source
Apache Mahout is a project developed by Apache Foundation that serves the primary purpose of creating machine learning algorithms. It focuses mainly on data clustering, classification, and collaborative filtering.
Mahout is written in JAVA and includes JAVA libraries to perform mathematical operations like linear algebra and statistics. Mahout is growing continuously as the algorithms implemented inside Apache Mahout are continuously growing. The algorithms of Mahout have implemented a level above Hadoop through mapping/reducing templates.

To key up, Mahout has following major features

- Extensible programming environment
- Pre-made algorithms
- Math experimentation environment
- GPU computes for performance improvement.

#9) Oracle Data Mining

Availability: Proprietary License
A component of Oracle Advance Analytics, Oracle data mining software provides excellent data mining algorithms for data classification, prediction, regression and specialized analytics that enables analysts to analyze insights, make better predictions, target best customers, identify cross-selling opportunities & detect fraud.

The algorithms designed inside ODM leverage the potential strengths of Oracle database. The data mining feature of SQL can dig data out of database tables, views, and schemas.

The GUI of Oracle data miner is an extended version of Oracle SQL Developer. It provides a facility of direct ‘drag & drop' of data inside the database to users thus giving better insight.

#10) Rattle

Availability: Open source
Rattle is GUI based data mining tool that uses R stats programming language. Rattle exposes the statistical power of R by providing considerable data mining functionality. Although Rattle has an extensive and well-developed UI, it has an inbuilt log code tab that generates duplicate code for any activity happening at GUI.

The data set generated by Rattle can be viewed as well as edited. Rattle gives the additional facility to review the code, use it for numerous purposes and extend the code without restriction.

#11) DataMelt

Availability: Open source

DataMelt, also known as DMelt is a computation and visualization environment that provides an interactive framework to do data analysis and visualization. It is designed mainly for engineers, scientists & students.

DMelt is written in JAVA and it is a multi-platform utility. It can run on any operating system which is compatible with JVM(Java Virtual Machine).

It contains Scientific & mathematical libraries.

Scientific libraries: To draw 2D/3D plots.
Mathematical libraries: To generate random numbers, curve fitting, algorithms etc.
DataMelt can be used for analysis of large data volumes, data mining, and stat analysis. It is widely used in the analysis of financial markets, natural sciences & engineering.

#12) IBM Cognos

Availability: Proprietary License
IBM Cognos BI is an intelligence suite owned by IBM for reporting and data analysis, score carding etc. It consists of sub-components that meet specific organizational requirements Cognos Connection, Query Studio, Report Studio, Analysis Studio, Event studio & Workspace Advance.

- **Cognos Connection:** A web portal to gather and summarize data in scoreboard/reports.
- **Query Studio:** Contains queries to format data & create diagrams.
- **Report Studio:** To generate management reports.
- **Analysis Studio:** To process large data volumes, understand & identify trends.
- **Event Studio:** Notification module to keep in sync with events.
- **Workspace Advanced:** User-friendly interface to create personalized & user-friendly documents.

#13) IBM SPSS Modeler

Availability: Proprietary License

IBM SPSS is a software suite owned by IBM that is used for data mining & text analytics to build predictive models. It was originally produced by SPSS Inc. and later on acquired by IBM.
SPSS Modeler has a visual interface that allows users to work with data mining algorithms without the need of programming. It eliminates the unnecessary complexities faced during data transformations and to make easy to use predictive models.

IBM SPSS comes in two editions, based on the features

- IBM SPSS Modeler Professional
- IBM SPSS Modeler Premium- contains additional features of text analytics, entity analytics etc.

#14) SAS Data Mining

Availability: Proprietary License
Statistical Analysis System (SAS) is a product of SAS Institute developed for analytics & data management. SAS can mine data, alter it, manage data from different sources and perform statistical analysis. It provides a graphical UI for non-technical users.

SAS data miner enables users to analyze big data and derives accurate insight to make timely decisions. SAS has a distributed memory processing architecture which is highly scalable. It is well suited for data mining, text mining & optimization.

#15) Teradata

Availability: Licensed
Teradata is often called Teradata database. It is an enterprise data warehouse that contains data management tools along with data mining software. It can be used for business analytics.

Teradata is used to have an insight of company data like sales, product placement, customer preferences etc. it can also differentiate between ‘hot' & ‘cold' data, which means that it puts less frequently used data in a slow storage section.

Teradata works on ‘share nothing' architecture as it has its server nodes have their own memory & processing ability.

#16) Board

Availability: Proprietary License
Board is often referred as Board toolkit. It is a software for Business Intelligence, analytics, and corporate performance management. It is a best-suited tool for companies looking to improve decision making. Board gathers data from all the sources and streamlines the data to generate reports in the preferred format.

Board is having most attractive and comprehensive interface among all BI software in the industry. Board provides facility to perform multi-dimensional analysis, control workflows and track performance planning.

#17) Dundas BI

Availability: Licensed
Dundas is another excellent dashboard, reporting & data analytics tool. Dundas is quite reliable with its rapid integrations & quick insights. It provides unlimited data transformation patterns with attractive tables, charts & graphs.

Dundas BI provides a fantastic feature of data accessibility from across many devices with a gap-free protection of documents.

Dundas BI puts data in well-defined structures in a specific manner in order to ease the processing for the user. It constitutes of relational methods that facilitate multi-dimensional analysis and focuses on business-critical matters. As it generates reliable reports, thus it reduces cost and eliminates the requirement of other additional software.