IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 20, Issue 5, Ver. II (Sep - Oct 2018), PP 12-19
www.iosrjournals.org

Analysis of Data Mining Tasks, Techniques, Tools, Applications And Trends

Dr.R. Shankar[1] and Dr.S. Duraisamy[2]
[1] (*Dept. of Computer Science, Chikkanna Government Arts College,Tirupur ,India)*
[2] (*Dept. of Computer Science, Chikkanna Government Arts College, Tirupur,India)*
Corresponding Author: R. Shankar

***Abstract*:** *Data mining is a process which finds useful patterns from huge amount of data. It is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. It uses machine learning, statistical and visualization techniques to discovery and present knowledge in a form which is easily comprehensible to humans. This paper deals with detail study of Data Mining its techniques, tasks and related Tools and also focuses on applications ad trends in the data mining which will helpful in the further research.*

Keywords: *Data mining, KDD, Clustering, Tools, Regression*.

Date of Submission: 23-09-2018 Date of acceptance: 08-10-2018

I. Introduction

Data mining is a hopeful and relatively new technology. Data mining is defined as a process of discovering hidden valuable knowledge by analyzing large amounts of data, which is stored in databases or data warehouse, using various core data mining techniques such as machine learning, artificial intelligence(AI) and statistical[1][2][3]. Figure 1.1 shows various data mining techniques. Many organizations in various industries are taking advantages of data mining including manufacturing, marketing, chemical, aerospace…etc, to increase their business efficiency. Therefore, the needs for a standard data mining process increased dramatically. A data mining process must be reliable and it must be repeatable by business people with little or no knowledge of data mining background.

Fig 1.1 Data Mining Techniques

- Sift through all the chaotic and repetitive noise in your data.
- Understand what is relevant and then make good use of that information to assess likely outcomes.
- Accelerate the pace of making informed decisions.

1.1 Need for Data Mining

- Data mining is the procedure of capturing large sets of data in order to identify the insights and visions of that data. Nowadays, the demand of data industry is rapidly growing which has also increased the demands for Data analysts and Data scientists.
- Since with this technique, analyze the data and then convert that data into meaningful information. This helps the business to take accurate and better decisions in an organization.

- Data mining helps to develop smart market decision, run accurate campaigns, predictions are taken and many more.
- With the help of Data mining, can analyze customer behaviors and their insights. This leads to great success and data-driven business. Figure 1.2 shows that the need for data mining[4].

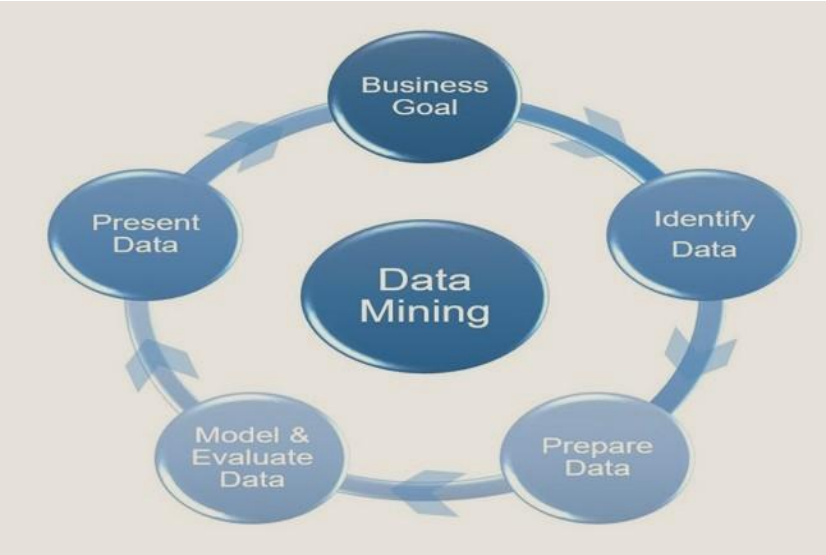

Fig 1.2 Need for Data Mining

1.2 Data Mining Tasks

1. **Create predictive power :** using features to predict unknown or future values of the same or other feature
2. C**reate a descriptive power:** find interesting, human-interpretable patterns that describe the data[5].

II. Data Mining Techniques

There are several major ***data mining techniques*** have been developing and using in data mining projects. The art of data mining has been constantly evolving[6]. There are several innovative and intuitive techniques that have emerged that fine-tune data mining concepts in a bid to give companies more comprehensive insight into their own data with useful future trends. Many techniques are employed by the data mining experts, some of which are listed below:

1. Classification:
This analysis is used to retrieve important and relevant information about data, and metadata. This data mining method helps to classify data in different classes.

2. Clustering:
Clustering analysis is a data mining technique to identify data that are like each other. This process helps to understand the differences and similarities between the data.

3. Regression:
Regression analysis is the data mining method of identifying and analyzing the relationship between variables. It is used to identify the likelihood of a specific variable, given the presence of other variables.

4. Association Rules:
This data mining technique helps to find the association between two or more Items. It discovers a hidden pattern in the data set.

5. Outer detection:
This type of data mining technique refers to observation of data items in the dataset which do not match an expected pattern or expected behavior. This technique can be used in a variety of domains, such as intrusion, detection, fraud or fault detection, etc. Outer detection is also called Outlier Analysis or Outlier mining.

6. Sequential Patterns:
This data mining technique helps to discover or identify similar patterns or trends in transaction data for certain period.

7. Prediction:
Prediction has used a combination of the other data mining techniques like trends, sequential patterns, clustering, classification, etc. It analyzes past events or instances in a right sequence for predicting a future event.

III. Data Mining Tools

Large amount of data generated every second and it is necessary to have knowledge of different tools that can be utilized to handle this large data and apply interesting data mining algorithms and visualizations in quick time[7].

i. Rapid Miner

It is one of the best predictive analysis systems. It provides an integrated environment for deep learning. The tool can be used for over a vast range of applications. As it includes for business applications, commercial applications, training, education, etc. 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. Also, it enables speedy delivery with a reduced number of errors.

ii. Orange

Orange is a perfect software suite for machine learning & data mining. It best aids the data visualization and is a component-based software.

iii. Weka

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.

iv. KNIME

KNIME is the best integration platform for data analytics. Also 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. It has been used for pharmaceutical research.

In addition, it performs for customer data analysis, financial data analysis. KNIME has some brilliant features like quick deployment and scaling efficiency. Users get familiar with KNIME in quite lesser time. Also, it has made predictive analysis accessible to even naive users.

v. Sisense

Sisense is extremely useful and best suited BI software. That it comes to reporting purposes within the organization. It has a brilliant capability to handle. Also, process data for the small-scale/large scale organizations. It allows combining data from various sources to build a common repository. Further, refines data to generate rich reports. That gets shared across departments for reporting. 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.

vi. SSDT (SQL Server Data Tools)

SSDT is a universal, declarative model. We use this model to expand all the phases of database development in the Visual Studio IDE. And developed to do data analysis and provide business intelligence solutions. Developers use SSDT transacts- a design capability of SQL and refactor databases. A user can work directly with a database. It can work with a connected database, thus, providing on or off-premise facility. Users can use visual studio tools for development of databases. Like IntelliSense, visual basic. SSDT provides Table Designer to create new tables. Also, edit tables in direct databases as well as connected databases.

vii. Apache Mahout

It 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. Such as linear algebra and statistics. Mahout is growing continuously as the algorithms implemented inside Apache Mahout. The algorithms of Mahout have implemented a level above Hadoop. Also. it is through mapping/reducing templates.

viii. Oracle Data Mining

A component of Oracle Advanced Analytics, it software provides excellent data mining algorithms. 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 a version of Oracle SQL Developer. It provides a facility of direct 'drag & drop' of data. That is inside the database to users thus giving better insight.

ix. Rattle

A rattle is a GUI 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. Also, it has an inbuilt log code tab that generates duplicate code for any activity happening at GUI.

x. DataMelt

DataMelt, also known as DMelt is a computation and visualization environment. Also, provides an interactive framework to do data analysis and visualization. It is designed mainly for engineers, scientists & students. DMelt is a multi-platform utility. It can run on any operating system which is compatible with JVM(Java Virtual Machine).

xi. IBM Cognos

IBM Cognos BI is an intelligence suite. It consists of sub-components that meet specific organizational requirements.

xii. IBM SPSS Modeler

IBM SPSS is a software suite owned by IBM. Also, we use it 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. Also, it allows users to work with data mining algorithm. Although, without the need for 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

xiii. SAS Data Mining

Statistical Analysis System (SAS) is a product of SAS Institute. SAS can mine data, alter it, manage data from different sources. Also, perform statistical analysis. It provides a graphical UI for non-technical users.SAS data miner enables users to analyze big data. And also 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.

xiv. Teradata

Teradata is often called Teradata database. It is an enterprise data warehouse. Also, it contains data management tools along with data mining software. We can use it for business analytics. Teradata works on 'share nothing' architecture. As it has its server nodes have their own memory & processing ability.

xv. Board

Board is often referred as Board toolkit. It is software for Business Intelligence, analytics, and corporate performance management. It is the best tool for companies looking to improve decision making. Board gathers data from all the sources. Also, streamlines the data to generate reports in the preferred format. Board is having most attractive and comprehensive interface. That it is among all BI software in the industry. Board provides facility to perform multi-dimensional analysis, control workflows and track performance planning.

xvi. Dundas BI

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. That is from across many devices with a gap-free protection of documents.

xvii. Python

As a free and open source language, **Python** is most often compared to R for ease of use. Many users find that they can start building data sets. And doing complex affinity analysis in minutes. The most common business-use case-data visualizations are straightforward. Although, till you are comfortable with basic programming concepts.

xviii. Spark

The attraction of **Spark** is plowing through vast oceans of data center traffic with ease. Park jobs run by Python. If you're moving into a big data, you'll need to know Spark. As it is one of the best open source data mining tools to deal with massive amounts of data.

xix. H2O

Want to get out on the cutting edge, start learning H2O. Also, it's been installed thousands of times, with applications for fraud detection. Like R, it has a very active and enthusiastic user community that's propelling its growth. Table 3.1 shows that various data mining tools and its importance of research .

Table 3.1 Data mining tools and its importance

S.No	Tool	Year and Author	Company/ Organization	Availability	Core Area	Focused on
1	Rapid Miner	2006, Ingo Mierswa and Ralf Klinkenberg	Technical University of Dortmund	Open Source	Data mining	Data science, machine learning, predictive analytics
2	Orange	1997	University of Ljubljana	Open Source	Data mining	Machine learning, Data mining, Data visualization, Data Analysis.
3	Weka	1993, University of Waikato	New Zealand	Free software	Data Mining, Python	Machine learning
4	KNIME	January 2004, KNIME.com AG	University of Konstanz	Open Source	Data mining	Enterprise Reporting / BI / Data Mining/ Deep Learning / Data Analysis / Text Mining
5	Sisense	2004, GuyBoyangu, Eldad Farkash, Adi Azaria, Aviad Harell	Tel Aviv, Israel, New York City, New York, United States	Licensed	Statistical, Business Intelligence	Business Discovery, Business Analysis, Software Company
6	SSDT	1989	Microsoft	Licensed	Visual Studio installation	Sql server data base, sql server analysis server
7	Apache Mahout	-	Apache software foundation	Open Source	machine learning algorithms, linear algebra and statistics, hadoop.	Data clustering, classification, and collaborative filtering.
8	Oracle Data Mining	1979	Oracle corporation	Proprietary License	Database maintenance	used for running OLTP, DW and mixed (OLTP & DW) database workloads
9	Rattle	2009 Dr.Graham Williams	Graham Williams organization	Open Source	Gui,R statistical package	File Inputs, Statistics ,Statistical tests, Clustering ,Modeling, Evaluation ,Charts, Transformations
10	DataMelt	S.Chekanov jHepWork **(20**05-2013) and **SCaVis** (2013-2015)	DataMelt community	Open Source	Data mining,java, python	DMelt creates high-quality vector-graphics images (SVG, EPS, PDF etc.) numeric computation, statistics, symbolic calculations, data analysis and data visualization.
11	IBM Cognos	2009, Alan Rushforth, Peter Glenister	Ottawa, Ontario, Canada	Proprietary License	statistical package	Business Intelligence, Operational Intelligence, Performance Management, Workforce Analytics
12	IBM SPSS Modeler	1968, Norman H. Nie, Dale H. Bent, and C. Hadlai Hull.	IBM Corporation, Armonk, New York	Proprietary License	statistical package	Statistical analysis, data mining, text analytics, data collection
13	SAS	1976, Anthony Barr	Sas insutitude	Proprietary License	analytics & data management,	Numerical analysis, statistical analysis
14	Teradata	1979, Victor L. Lund CEO, Kenneth Simonds,	San Diego, California,	Licensed	database and analytics-related products and	business analytics, cloud products, and consulting

					services	
15	Board	1994, BOARD International	Chiasso, Switzerl and,	Proprietary License	perform multi-dimensional analysis, control workflows	create and update databases, data presentations, analyses and process models, in a single visual and interactive environment
16	Dundas BI	1992 ,Microsoft	Toronto, Ontario,Canada, Microsoft	Licensed	dashboard, reporting & data analytics tool	Dundas BI, Dundas Dashboard, Dundas Consulting Professional Services
17	Python	1990, Guido van Rossum	Python software foundation	Open Source	Machine learning	Python is great for Web Development, complex data analysis and visualization.
18	Spark	2014, Matei Zaharia	Apache Software Foundation, UC Berkeley AMPLab , Databricks	Open Source	Analytics engine and Big data	cluster manager and a distributed storage system, Hadoop YARN, or Apache Mesos,
19	H20	2011,h20 ai	H2o , us california	Open Source	Big data analysis, python	machine learning, statistical learning theory
20	Tanagr a	2003, France	Lumière University Lyon 2	Open Source	data analysis, statistical learning, machine learning and databases area.	some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and non parametric statistics, association rule, feature selection and construction algorithms
21	R	1993,Ross Ihaka and Robert Gentleman	University of Auckland, New Zealand, and is currently developed by the R Development Core Team.	Open Source	statistical computing	linear and nonlinear modeling, classical statistical tests, time-series analysis, classification, clustering, and others. R is easily extensible through functions .
22	XL Miner	-	Microsoft	Licensed	statistics, machine learning, and artificial intelligence,	Data Analysis, Time Series, Data Mining

IV. Data Mining Applications

Data mining applications in various areas including sales/marketing, banking, insurance, healthcare, transportation, and medicine[8][9].

Data Mining Applications in Sales/Marketing

Data mining enables businesses to understand the hidden patterns inside historical purchasing transaction data, thus helping in planning and launching new marketing campaigns in a prompt and cost-effective way. The following illustrates several data mining applications in sale and marketing.

- Data mining is used for market basket analysis to provide information on what product combinations were purchased together when they were bought and in what sequence. This information helps businesses promote their most profitable products and maximize the profit. In addition, it encourages customers to purchase related products that they may have been missed or overlooked.
- Retail companies use data mining to identify customer's behavior buying patterns.

Data Mining Applications in Banking / Finance

- Several data mining techniques e.g., distributed data mining have been researched, modeled and developed to help credit card fraud detection.
- Data mining is used to identify customers loyalty by analyzing the data of customer's purchasing activities such as the data of frequency of purchase in a period of time, a total monetary value of all purchases and when was the last purchase. After analyzing those dimensions, the relative measure is generated for each customer. The higher of the score, the more relative loyal the customer is.
- To help the bank to retain credit card customers, data mining is applied. By analyzing the past data, data mining can help banks predict customers that likely to change their credit card affiliation so they can plan and launch different special offers to retain those customers.

- Credit card spending by customer groups can be identified by using data mining.
- The hidden correlation's between different financial indicators can be discovered by using data mining.
- From historical market data, data mining enables to identify stock trading rules.

Data Mining Applications in Health Care and Insurance

The growth of the insurance industry entirely depends on the ability to convert data into the knowledge, information or intelligence about customers, competitors, and its markets.

Data mining is applied in insurance industry lately but brought tremendous competitive advantages to the companies who have implemented it successfully. The data mining applications in the insurance industry are listed below:

- Data mining is applied in claims analysis such as identifying which medical procedures are claimed together.
- Data mining enables to forecasts which customers will potentially purchase new policies.
- Data mining allows insurance companies to detect risky customers' behavior patterns.
- Data mining helps detect fraudulent behavior.

Data Mining Applications in Transportation

- Data mining helps determine the distribution schedules among warehouses and outlets and analyze loading patterns.

Data Mining Applications in Medicine

- Data mining enables to characterize patient activities to see incoming office visits.
- Data mining helps identify the patterns of successful medical therapies for different illnesses.

Data mining applications are continuously developing in various industries to provide more hidden knowledge that increases business efficiency and grows businesses.

V. Data Mining Trends

Data mining concepts are still evolving and here are the latest trends that we get to see in this field –[10][11]

- Application Exploration.
- Scalable and interactive data mining methods.
- Integration of data mining with database systems, data warehouse systems and web database systems.
- Standardization of data mining query language.
- Visual data mining.
- New methods for mining complex types of data.
- Biological data mining.
- Data mining and software engineering.
- Web mining.
- Distributed data mining.
- Real time data mining.
- Multi database data mining.
- Privacy protection and information security in data mining.

VI. Conclusion

In this paper we have discussed the detail study of various data mining tasks, techniques, tools, applications and trends. This review would be helpful to researchers to focus on the various issues of data mining. 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. These tools predict future trends and behaviors, allowing business to make proactive and present knowledge in the form which is easily understood to human.

References

[1]. M H Dunham, *Data Mining: Introductory and Advanced Topics*, Prentice Hall, 2002.
[2]. Dunham, M. H., Sridhar S, *Data Mining: Introductory and Advanced Topics*, Pearson Education, New Delhi, ISBN: 81-7758-785- 4, 1st Edition, 2006
[3]. http://www.zentut.com/data-mining/data-mining-processes/
[4]. https://www.loginworks.com/blogs/217-data-mining-and-its-importance/
[5]. https://blog.galvanize.com/four-data-mining-techniques-for-businesses-that-everyone- should-know/
[6]. https://www.guru99.com/data-mining-tutorial.html
[7]. https://www.softwaretestinghelp.com/data-mining-tools/
[8]. http://www.zentut.com/data-mining/data-mining-applications/

[9]. https://www.tutorialspoint.com/data_mining/dm_applications_trends.htm
[10]. Article Practical Applications of Data Mining: Trends in Data Mining Web Source http://www.dataminingtools.ne
[11]. Krzysztof J. Cios and Lukasz A. Kurgan Trends in *Data Mining and Knowledge Discovery* Web Source: isds.bus.lsu.edu

IOSR Journal of Computer Engineering (IOSR-JCE) is UGC approved Journal with Sl. No. 5019, Journal no. 49102.

R. Shankar. " Analysis Of Data Mining Tasks, Techniques, Tools, Applications And Trends." IOSR Journal of Computer Engineering (IOSR-JCE) 20.5 (2018): 12-19.