ISSN No:-2456-2165

Multidimensional Data Analysis, Data Mining and Knowledge Discovery

Ekwe Prince O.[1, 2]; Okoronkwo Mathew[1]; Ukwome Tochi P[2]; Anozie Valentine U[2]
[1]Department of Computer Science, University of Nigeria, Nsukka (UNN)
[2]Department of Computer Science, Federal College of Agriculture, Ishiagu (FCAI)

Abstract:- In recent times, the rate of usage and consumption of data has led to the need for these data to be organized, analyzed and used for futuristic prediction and decision making in order to improve human lives and future prediction in different fields of endeavor. Multidimensional Data Analysis, Data Mining and Knowledge Discovery are all associated with the organization, analysis and extraction of a data set for organization's decision making and futuristic prediction. In this research work, our focus was on the techniques, application of data mining as well as the phases involved in data mining. Our work further highlighted the current and future trends in data mining and the numerous positives of multidimensional data analysis/data mining and knowledge discovery to individuals, organizations, government, societies and the world at large. The outcome of this research would provide a detailed positive view of the impact of data mining to individuals, organizations, government, societies and the world at large for decision making and futuristic prediction.

Keywords:- *Data, Mining, Data Mining, Multidimensional Data Analysis, Knowledge Data Discovery*.

I. INTRODUCTION

The advent and wide spread of information technology has led to continuous generation of data in different databases. These generated data set needs to be organized and extracted with the sole aim of fine tuning useful intelligence to help organizations/businesses resolve imminent challenges, predict trends, mitigate risk and find trending opportunities to improve human lives.

We live in an information-rich world which is data-driven. It is awesomely comforting to wake up to surplus of readily accessible data, information and knowledge, these large amount of available data gives rise to multiple challenges, organization needs and extraction of useful information for decision making. The Information depth would determine the useful insights you need.

Data mining takes advantage of big data's endless opportunities and affordable processing power to analyze large data with ease. Computer delivery power and speed have improved substantially in recent times, which has allowed the whole wide world to experience rapid, easy, and automated data analysis.

Data mining involves the extraction of meaningful information and patterns from complex data for the sake of decision making and futuristic prediction. It is also referred to as knowledge discovery process, knowledge mining data, knowledge extraction, multidimensional data analysis or data /pattern analysis.

Data mining is also used for establishing connection and creating variant patterns, abnormalities and correlations used for tackling multiple issues as well as processing usable information in the process. Data discovery is a vast, varied process that involves multiple elements which gives birth to refined decisions.

The remaining parts of these work is organized and arranged into section two which highlights the background of data mining, section three lays emphases on the phases of data mining while section four sheds light on the application areas of data mining. Section five educates on the benefits of data mining to organizations and section six points out the different techniques of data mining. Section seven breaks down the current and future trends of data mining, while section eight highlights the contributions to knowledge and section nine concludes the research.

II. BACKGROUND TO THE STUDY

Angeli et al., (2017) asserted that for millennia now, humans have dug out different locations to search for missing knowledge. "Knowledge discovery in databases" is the process of sieving different data to find hidden information and forecast varying future trends. The phrase "data mining" came into being in the 1990s. Data mining came to light from the convergence of three scientific disciplines, these disciplines are: artificial intelligence, machine learning, and statistics.

The 21st century presents us with high-dimensional, large-scale, distributed digital mining in which enormous data are been mined in short period of time leading to bright-line prospects, and the potential positive value is also boundless. Among them, the classification prediction technology will aid in future smart economic activities as well as provide vital reference decisions.

Li and Long (2020) researched on image detection and quantitative detection analysis of gastrointestinal infections using data mining. Zuo (2018) did a detailed analysis on the attributes of network viruses and developed an electronic data mining software. He further blended the data mining

technology and dynamic behavior interception technology to mine encrypted data and ascertain whether there is a malicious program. This approach was employed to network Trojan virus discovery.

Although alternating analytic methods are preferred to be used for varying data sources with particular characteristics, some prominent analytic methods can be carried out centered on the common characteristics of log files.

Hao et al. (2016) summarized a number of common actions when launching the package in Python, glassPy. These includes the summary data from the log file, the number of sessions, the time duration of each session and the frequency of each event. More so, event n-grams, or event sequences of different lengths can be created for further implementation of similarity measures to classify and rate individuals' performances. To take into account, the temporal data, hierarchical vectorization of the rank ordered time intervals and the time interval distribution of event pairs were also implemented.

In addition to these common electronic data analytic techniques, other existing data analytic approaches for processing data are Social Network Analysis (SNA; Zhu et al., 2016), Bayesian Networks/Bayes nets (BNs; Levy, 2014) and Markov Item Response Theory (Shu et al., 2017). Furthermore, later data mining approaches, including cluster analysis, decision trees, and artificial neural networks have been useful in unveiling vital information about students' problem-solving strategies in different technology-improved grading.

Buczak and Guven (2017) produced a hands on lecture on machine learning (ML) approaches and data mining (DM) processes for network analysis. Xu et al. (2013) explored the intermediate problems associated to data mining from a wider horizon and detailed several approaches that aid in preserving sensitive information. He reviewed recent and trending approaches of Data Mining and came up with some preliminary nuggets for futuristic research. Yan and Zheng (2017) discovered that long after doing a detailed job on data mining, many basic signs are vital predictors of cross-sectional stock benefits. Their approaches are general and it was used on past benefit-based anomalies. Emoto et al. (2017) used terminal restriction fragment length polymorphism (T-RFLP) data mining technology to show the gut microbiota profile of patients who have coronary artery disease. Hong et al. (2018) presented a modern approach to construct a flood sensitivity map in Poyang County, Jiangxi Province, China, by implementing the fuzzy Wolfe and data mining procedures. The data output of these studies are not broad-gauged and the outputs lack footing; thus, they cannot be completely recognized by the nuetrals.

This study aims to highlight data mining trends, techniques and application, this paragraph provides a brief review of related techniques that have been frequently used and the impact of these researches in relation to analyzing process data in technology-driven system. There are two (2) main classes of data mining techniques, they are supervised and unsupervised learning methods (Fu et al, 2014; Sinharay, 2016). Supervised approaches are used when subjects' memberships are known and the reason is to train a classifier that can concisely classify the subjects into their own category (e.g., score) and then be efficiently generalized to new datasets. Unsupervised approaches are used when subjects' memberships are unknown and the goal is to categorize the subjects into clearly different clusters based on characteristics that can differentiate them. Decision tree is a supervised data classification approach has been utilized very often in analyzing process data in varying systems.

DiCerbo and Kidwai (2013) worked with Classification and Regression Tree (CART) methods to produce classifiers to detect a player's goal in a gaming surrounding. The authors showed the creation of the classifier including feature generation, pruning process, and evaluated the results using concise data. This research showed that the CART could be a dependable automated detector and showed the procedure of how to create such a detector with a relatively small sample size (n = 527).

On the other hand, cluster analysis and Self-Organizing Maps (SOMs) are two pronounced unsupervised methods that organize students' problem-solving strategies. These shows that cluster analysis can constantly identify key characteristics in 155 students' performances in log files extracted from an educational gaming and simulation surrounding called *Save Patch*, which measures mathematical competence. The authors showed how they manipulated the data for the application of clustering algorithms and identified evidence that fuzzy cluster analysis is more accurate than hard cluster analysis in analyzing log file process data from game/simulation surrounding. Most importantly, the authors showed that cluster analysis can identify both effective strategies and misconceptions students have with respect to the related construct. Fossey (2017) reviewed three unsupervised approaches, including *k*-means, SOM and Robust Clustering using Links (ROCK) on analyzing process data in log files from a game-based assessment case.

III. STEPS INVOLVE IN DATA MINING

Data Mining has some steps involve in actualizing analysis of varying data in order to organize and analyze data for decision making and futuristic prediction.

Data mining provides an indebt knowledge of arithmetic/statistics, programming, business principles as well as communication. In gathering knowledge about data analysis, all data scientist must note: Linear Algebra, Machine Learning, Data Retrieval and Database, Artificial Intelligence, Problem-solving Ability, Data Structures and Algorithms, Statistical Analysis.

➢ *Outlined below are the Processes Data Analysts/Scientists Follow in Order to Handle Data Mining Project;*

- *Broad Understanding of the Business/organization: At this stage of Data Mining, the following data and information must be studied and understood in detail. This information include;*

✓ *Basic Knowledge of the Company*
✓ *The Organization's Present Footing,*
✓ *The Project's Goals/Objectives,*
✓ *What is the Benchmark for Success?*

- *Comprehend the type of Data: This stage involves a thorough understanding of the data to be analyzed and the various sources of the data. In more simpler terms, there are two steps involve in this stage;*

✓ Decipher the kind of data that is needed to resolve the problem
✓ Source for data from the right source.

- *Arrange the Data: These stage of Data Mining involves sorting, arranging and preparing data for analyzing. The steps in this phase are outlined below;*

✓ Ramify data quality issues such as duplicate data, missing data or infected data
✓ Arrange the data in a structure acceptable to solve the organization's challenges.

- *Remodel the Data to a particular form: This stage of Data Mining entails employing algorithm to ascertain data patterns after which a working model is created. The steps in this stage are outlined below;*

✓ Deploy algorithms to predict data patterns.
✓ Data analyst create, test, and ascertain a suitable model from data patterns generated.

- *Evaluate the Data to ascertain the result: This stage decides whether/how effective the outcome of the model will positively impact the business goal or solve the issue. More so, a repetitive level for sourcing the appropriate algorithm is proposed, in a situation where the data analyst fails to attain success in the first instance.*
- *Deploy the System to Management: The outcome of data mined is given to the management for decision making and futuristic prediction.*

IV. APPLICATION AREAS OF DATA MINING

The positives of Data mining in competitive businesses environment is enormous, as a result, decision making and future prediction is achieved in a short period of time. Outlined below are some data mining examples that shows a broad range of application areas.

A. The Following are Trends and Application Areas of Data Mining:

➢ *Shopping Data Analysis*

The shopping market present us with a large data, the management may need to swim through this large chunk of data by different patterns. In order to achieve this fit, market basket analysis is an appropriate analytic method. Market basket analysis is an analytic approach that uses the idea that once you buy one set of goods, you're likely to buy another set of goods. This approach helps small scale businesses predict a customer's purchasing habits. With differential analysis, data gotten from various customers and clients from various regional clusters can be analyzed.

➢ *Weather Prediction Analysis*

Data Mining is also used for weather prediction, weather forecasting systems makes decision from large amounts of historical data for a particular period of time. Since large chunks of data are accessed and processed, the right data mining approach would be deployed.

➢ *Analysis in the Stock Market*

The stock market deals with enormous chunk of data, these data needs to be analyzed. As a result, data mining approaches are used to model such data in order to perform the analysis.

➢ *Intrusion Detection System*

Intrusion Detection System analysis varying data in order to detect the network activity. Data mining enhances intrusion detection system by predicting anomaly detection. It helps in differentiating between unusual network activity and normal network activity.

➢ *Fraud Detection System*

Data Mining aids in Fraud Detection System. Old fashion approaches of fraud detection are time-wasting and stressful as a result of the massive chunk of data involved. Data mining helps in predicting relevant patterns and the processing of data into information.

➢ *Video Surveillance*

Video surveillance is used practically everywhere in our everyday life for security reasons. Enormous Data are captured every single day from the cameras and we need Data Mining for analyzing the enormous chunks of data.

➢ *Analysis in Financial Banking*

The importance of Data Mining in Financial Banking cannot be underestimated. Because for every fresh business deal in automated banking, an enormous chunk of data is produced. By resolving hidden patterns, causalities, and correlations in business data, data mining would definitely provide solutions in banking bottlenecks and data access in banking and finance.

Data mining aids financial institutions ascertain anti-fraud systems and credit ratings, analyze client financial data, transaction record and financial card purchases. More so, financial institutions get a better understanding of their

clients' online habits and preferences through Data Mining, which serves as a bedrock when creating a new marketing campaign.

➢ *Data Mining Analysis in Healthcare*

Medical Practitioners create precise diagnosis by combining physical examination results, patient's medical history, medications, and treatment patterns using Data Mining. It also reduces fraud and waste as well as welcome a more cost-effective automated health resource management system.

➢ *Data Analysis in Marketing*

Marketing is one of the application that benefits most from data mining, it is actually marketing! After all, the sole aim of marketing is to ascertain clients effectively for utmost sales and the most appropriate strategy to get your customers is to know your customers. Data mining comes in handy in bringing as a unit data on gender, income level, age, location, tastes and spending habits to produce satisfying individualized loyalty campaigns. Data Analysis can also ascertain which clients will most possibly unsubscribe to a mailing list or other service offered by the organization. With such data from Data Mining techniques, businesses are sure to take appreciable strategies to ensure such clients don't unsubscribe and leave for other competitors.

➢ *Data Analysis in Retail*

Data mining in retail and marketing work in parallel, but the former still needs its personal highlight. Retail shops and supermarkets uses client choice patterns to clamp down commodity associations and predict the items that need to be stocked in the supermarket/business premises and where it can be gotten. Data mining also highlights which of the choices attain the most response.

V. BENEFITS DERIVED FROM DATA MINING

We have our being in a data-driven globe, this gives us a lot of advantages with ease of data access. Data mining gives us a paradigm for solving and resolving challenges in this complex information age. The benefits of Data mining include:

- It aids in gathering reliable information for organization/businesses
- It is efficient and cost less in comparison to most data systems around
- It aids varying companies to ascertain all round profit and model shifts
- Data mining can work with both new and legacy computer systems
- Informed decisions are made through data mining for businesses and organization
- It also aims at detecting various credit risks and fraud
- Scientists analyze large amounts of data with ease using Data Mining
- Data analyst use the information to detect fraud, create risk models, and increase product safety
- It aids data analyst to quickly instantiate computerized predictions of behaviors and trends and discover hidden patterns

VI. DATA MINING TECHNIQUES/TOOLS

As seasoned work men are known for the saying, "To achieve success, use the right tool for the right job." It is very vital to note the different tools used for data mining in order to make accurate and prompt organizational decision. Outlined below are the strategies/techniques that aid data scientists with multiple data mining abilities.

➢ *Artificial Intelligence Tool*

Artificially Intelligent process produce analytical functions that reproduce human intelligence, such as problem-solving, reasoning, planning, and learning.

➢ *Association Rule Learning Technique*

The association rule toolset is also known as market basket analysis, it searches for relationships amongst dataset variables. A case study is association rule learning can determine which commodities are always bought together (e.g., a smartphone and a protective case).

➢ *Clustering Technique*

The clustering technique organizes datasets into different useful sets, known as clusters. This technique aids individuals perceive the normal bedrock or strata in the data.

➢ *Classification Technique*

The classification technique allocates particular data in a set to various particular clusters or segments. The main objectives is to create concise forecasts in the particular cluster for all the members of the set.

➢ *Data Analytics Technique*

The data analytics technique allows users to ascertain digital information and create useful business intelligence from it.

➢ *Data Cleansing and Preparation Technique*

This Technique processes the raw data to an optimal pattern suitable for later processing and usage. Preparation involves processes such as finding and debugging errors and checking for missing or duplicate data.

➢ *Data Warehousing Technique*

This technique includes a comprehensive collation of company's data that management utilize to aid them in setting appropriate goals. Warehousing is a basic and important jig saw of major large data mining systems.

➢ *Machine Learning Technique*

Machine Learning Technique is related to the Artificial Intelligent technique, machine learning is a computerized programming technique that uses statistical probabilities to create computers with the skills to learn without human intervention or manually programmed.

➢ *Regression Technique*

The regression technique forecasts a number of numeric data in clusters such as sales, stock prices, or even temperature. The organization are based on the information found in a particular data set.

- *There are Two Specific Tools, they are: R and ODM*

R is an open source language that is used for graphics and statistical computing. It enrich data scientist with a wide option of statistical tests, classification and graphical techniques, and time-series analysis.

Oracle Data Mining (ODM) is a module of the Oracle Advanced Analytics Database. It aids data scientists in forecasting and creating ambiguous insights. Data Scientist use ODM to forecast client behavior, develop client profiles, and identify cross-selling opportunities.

VII. CURRENT AND FUTURE TRENDS IN DATA MINING

The current and hereafter of data mining is positive, as long as data volumes continually increase. Data mining techniques have improved due to technology improvement, as have systems that extract useful information from data improved also. Before now, only large organizations with high budgets utilized multiple supercomputers to ascertain data for organizing, analyzing and mining data for decision making and futuristic prediction due to the expenses of saving and analyzing data was very cost.

Today, Organizations are introducing artificial intelligence, machine learning, and deep learning on cloud-based data lakes to extract vital information from chunks of data for futuristic prediction.

The Internet of Things and wearable computing has changed both individuals and electronic devices into data-generating machines capable of creating massive data on individuals and organizations. Through this, organizations can accumulate, store, and analyze massive amounts of data for decision making.

Cloud-based analytics solutions provide an easier and more cost-effective paradigm for organizations to entertain large amounts of data and processing power. Cloud computing allows for organizations to easily access and react on data from manufacturing, sales, and inventory systems, Internet, marketing, among other sources in order to enhance their bottom line.

VIII. CONTRIBUTION TO KNOWLEDGE

This research paper tends to contribute to knowledge by highlighting the place of organizing and extracting information from a data set for decision making and futuristic prediction. More so, the research paper tends to bring to lamplight, the techniques involve in data mining and the current and future trends for individuals, organizations, government, societies and the world at large.

IX. CONCLUSION

Data Mining, Multidimensional Data Analysis and Knowledge Discovery are integral parts of cumbersome data analysis, organization, extraction and presenting needed information for decision making and futuristic prediction. Needless to say that the role and importance of data mining to individuals, organizations, government, societies and the world at large cannot be undermined.

The article provided a comprehensive knowledge on where Data Mining started, it did go one step further to explore the trends and techniques of data mining. Finally, the future and applications of Data Mining were also exhausted.

REFERENCES

[1]. Angeli, C., Howard, S. K., Ma, J., Yang, J., and Kirschner, P. A. (2017). Data mining in educational technology classroom research: can it make a contribution? *Computers & Education*, vol. 113, pp. 226–242.

[2]. Buczak A. and Guven. (2017). A survey of data mining and machine learning methods for cyber security intrusion detection, *IEEE Communications Surveys & Tutorials*, vol. 18, no. 2, pp. 1153–1176.

[3]. DiCerbo, K. E., and Kidwai, K. (2013). Detecting player goals from game log files, in *Poster presented at the Sixth International Conference on Educational Data Mining* (Memphis, TN).

[4]. Emoto, T., Yamashita, T., Kobayashi, T. (2017). Characterization of gut microbiota profiles in coronary artery disease patients using data mining analysis of terminal restriction fragment length polymorphism: gut microbiota could be a diagnostic marker of coronary artery disease, *Heart and Vessels*, vol. 32, no. 1, pp. 39–46.

[5]. Fossey, W. A. (2017). An Evaluation of Clustering Algorithms for Modeling Game-Based Assessment Work Processes. *Unpublished doctoral dissertation, University of Maryland, College Park*.

[6]. Fu, J., Zapata-Rivera, D., and Mavronikolas, E. (2014). Statistical Methods for Assessments in Simulations and Serious Games (ETS Research Report Series No. RR-14-12). Princeton, NJ: Educational Testing Service.

[7]. Hao, J., Smith, L., Mislevy, R. J., von Davier, A. A., and Bauer, M. (2016). Taming Log Files From Game/Simulation-Based Assessments: *Data Models and Data Analysis Tools* (ETS Research Report Series No. RR-16-10). Princeton, NJ: Educational Testing Service.

[8]. Hong, H., Tsangaratos, P., Ilia, I., Liu, J., Zhu, A.-X. and Chen, W. (2018). Application of fuzzy weight of evidence and data mining techniques in construction of flood susceptibility map of Poyang County, China," *The Science of the Total Environment*, vol. 625, no. 1, pp. 575–588.

[9]. Levy, R. (2014). Dynamic Bayesian Network Modeling of Game Based Diagnostic Assessments (CRESST Report No.837). Los Angeles, CA: *University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Center for Studies in Education, UCLA*.
[10]. Li, T. and Long, L. (2020). Imaging examination and quantitative detection and analysis of gastrointestinal diseases based on data mining technology, *Journal of Medical Systems*, vol. 44, no. 1, pp. 1–15.
[11]. Sinharay, S. (2016). An NCME instructional module on data mining methods for classification and regression. *Educ. Meas. Issues Pract.* 35, 38–54. doi: 10.1111/emip.12115.
[12]. Shu, Z., Bergner, Y., Zhu, M., Hao, J., and von Davier, A. A. (2017). An item response theory analysis of problem-solving processes in scenario-based tasks. *Psychol. Test Assess. Model.* 59, 109–131.
[13]. Xu, B., Recker, M., Qi, X., Flann, N., and Ye, L. (2013). Clustering educational digital library usage data: a comparison of latent class analysis and k-means algorithms. *J. Educ. Data Mining* 5, 38–68.
[14]. Yan, X. and Zheng, L. (2017). Fundamental analysis and the cross-section of stock returns: a data-mining approach, *Review of Financial Studies*, vol. 30, no. 4, pp. 1382–1423.
[15]. Zhu, M., Shu, Z., and von Davier, A. A. (2016). Using networks to visualize and analyze process data for educational assessment. *J. Educ. Meas.* 53, 190–211. doi: 10.1111/jedm.12107.
[16]. Zuo, C. (2018). Defense of computer network viruses based on data mining technology, *International Journal on Network Security*, vol. 20, no. 4, pp. 805–810.