ISSN No:-2456-2165

A Brief Survey: Data Mining Techniques and Application on Selected Sectors

Grace Etiowo Jackson
ICT Department
NABTEB
Benin City, Nigeria.

M. C. Okoronkwo
Department of Computer science
University of Nigeria,
Nsukka, Nigeria

Abstract:- Data mining is described as a systematic process that is a step in the process of information discovery and is based on algorithms to produce patterns and data analysis applications. This paper discusses some of the data mining techniques and its applications in some selected fields of human endeavors.

Keywords *- Data Mining, Knowledge Discovery Process, Data Mining Techniques, Data Mining Applications,*

I. INTRODUCTION

We live in a time that is commonly tagged as the technological era. In this era of information, because we think that knowledge gives us power and prosperity and because of advanced technology like computers, we have been gathering a ton of data via satellites and other technologies. When computers first became available, along with methods for mass digital storage, humans began gathering and storing a wide variety of data, relying on the computing power of computers to help us filter through this deluge of data. Unfortunately, these vast databases of information housed on several platforms quickly became too much to handle. Structured databases and database management systems (DBMS) were developed as a result of the original turmoil [1]. For managing a big body of data, effective database management solutions have proved a huge advantage and notably for effective and speedy retrieval of specific data whenever needed from a huge collection. The current huge collection of information of all kinds has also been facilitated by the growth of database management systems. We now have significantly more information than we can handle, including data from scientific research, corporate transactions, satellite images, written reports, and military intelligence [1].

Simply said, gathering information is no longer sufficient for decision-making. Now that we are faced with massive data sets, we have developed new demands to guide our management decisions [2]. These requirements include the automated summary of data, the extraction of the "essence" of the stored information, and the identification of patterns in the raw data.

Data mining is a formidable new technology that has the ability to greatly assist businesses in concentrating on the most crucial data in their data warehouses. It is described as: “The computerized examination of extensive or complicated data sets in order to identify important patterns or trends that might otherwise go unnoticed”

II. CONCEPTS AND TECHNIQUES

A. Overview of Data mining and Knowledge Discovery

Data mining, in the words of Kleinberg and his coworkers [6] is "an interesting pattern extraction process from the raw data". Data mining is the act of "discovering new correlations, patterns and trends that are meaningful by passing a large number of data stored in a vault," according to the Gartner Group [7]. Another definition of data mining is the practice of obtaining important information from the vast amounts of data gathered by businesses [9].

Data mining has developed to find usable data in databases, eliminate duplicate data, and provide correct data as quickly as feasible. In order to uncover some future predictions, data mining makes it simpler to search for a set of rules inside a vast quantity of data. It also aids in the extraction and utilization of useful data from vast amounts of data.

Data mining is the core of knowledge Discovery Process. The Knowledge Discovery Process comprises of a few steps leading from raw data collections to some form of new knowledge. The steps involve in this iterative process are:

- Data cleaning: at this stage, the noisy, erroneous, and inconsistent data in the data is removed and the missing data is completed, it is also referred to as data cleansing.
- Data integration: here, a plurality of data sources from different sources are consistently combined in one source;
- Data selection: at this step, the data relevant to the analysis to be performed is determined and taken from the database.
- Data transformation: also known as data consolidation, it is a phase where the data is prepared for the data analysis methods to be used.
- Data mining: it is the crucial step in which intelligent methods are used to extract data patterns.
- Pattern evaluation: this step establishes the correct and interesting patterns representing the information obtained in accordance with the measurements made.
- Knowledge representation: this is the final stage, which involves presenting the information to the user using information exploration techniques.

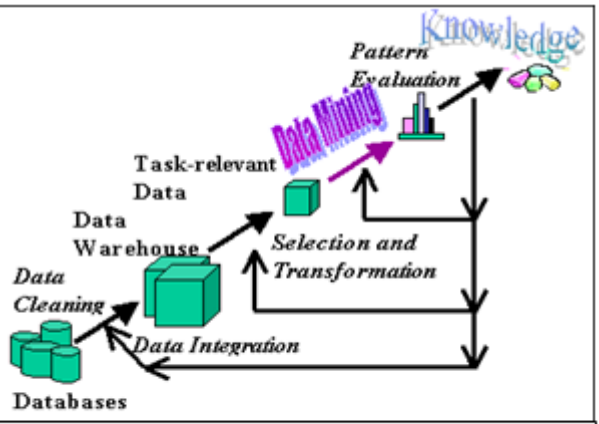

Figure 1: Knowledge Discovery Process

B. Data Mining Techniques

Algorithms and other methods are used in data mining to transform massive data sets into useable output. The most often used kinds of data mining methods are as follows:

➢ *Association rules:*

The association rules look for correlations between variables, commonly known as market basket analysis. As it attempts to connect different bits of data, this relationship in and of itself adds value to the data collection. For instance, association rules would examine a business's sales data to determine which goods were most frequently bought in a particular company; with this knowledge, retailers might plan, advertise, and forecast. There are different types of association rules namely:

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

➢ *Classification:*

This is the most commonly applied data mining technique. It assigns pre-defined classes to objects. These categories express the traits of the items or the similarities between the data points. The underlying data may be more precisely classified and summed up across related attributes or product lines thanks to this data mining approach [3]. There are different types of classification model namely:

- Classification By Decision Tree
- Bayesian Classification
- Neural Networks
- Support Vehicle Machines (SVM)
- Classification Based On Associations

➢ *Clustering:*

Clustering is comparable to categorization. Clustering, on the other hand, finds commonalities between things before classifying them according to how they vary from one another. There are different types of clustering methods namely:

- Partition Methods
- Hierarchical Agglomerative (divisive) method,
- Density Based Methods
- Grid Based Methods
- Model-based Method

➢ *Decision Trees:*

Decision trees are used to categorize or forecast a result based on a predetermined set of standards or judgments. In order to rank the dataset depending on the answers provided, a decision tree is utilized as the input for a series of cascading questions[4]. A decision tree provides for particular direction and user input when digging deeper into the data and is occasionally represented visually as a tree.

➢ *K-Nearest neighbor (KNN):*

This is an algorithm that sorts data according on how close it is to other data. The underlying premise of KNN is that data points that are near to one another have a higher degree of similarity than other types of data [4] This supervised, non-parametric method predicts the characteristics of a group from a set of individual data points.

➢ *Neural Networks:*

In this technique, data processing nodes are used. These nodes consist of an output, weights, and inputs. Data is mapped using supervised learning, which is comparable to how the human brain is wired. The accuracy of a model may be determined by programming this model to provide threshold values.

➢ *Predictive analysis:*

Predictive analysis attempts to use previous data to construct graphical or mathematical models that can predict future results. Similar to regression analysis, this approach seeks to support an unknown figure in the future using information already available.

C. Data Mining Application in selected sectors

To obtain a competitive edge and support corporate expansion, many companies employ data mining to crucial business operations.

Numerous industries, including banking, the stock market, marketing, retail sales, signal processing, insurance, telecommunication, health, medicine, biology, genetics, industry, construction, education, intelligence, science, and engineering are among the many fields where data mining has applications [5, 10, 11]. Application of data mining in the major industries that encompass the aforementioned industries are considered in this paper , and they were described with examples of how and why data mining was employed in those industries.

➢ *Education Sector*

Data mining has been used in numerous ways in the field of education which include:

- students' performance analysis, determining the status of students' pass and fail;
- determining factors affecting the success of students enrolling in higher school;
- creating the preference of university departments;
- determining the factors that influence the preference order of new enrolled students (18);

- choosing a profession based on the demographic and personal characteristics [14],
- Determine the correlations between the type of school from which students graduate and their university departments [15, 16], evaluate the study habits of distance education students [16, 17], prevent students from failing, and identify the elements that impact success.
- Finding out about the profiles and preferences of students taking the university entrance exam [18], figuring out the connection between academic success and extracurricular activity participation among university students, figuring out the socioeconomic status of students and their level of academic learning [14], and figuring out whether there is a link between entry test scores and academic success.

These applications in the educational sector support instructors by assisting them in organizing their lessons, comprehending student learning, and giving students' pro-active feedback.

➢ *Banking and Finance Sector*

Data mining is widely used in the banking and finance sector to ascertain what, when, and why the customer profile chooses. The search for suitable solutions for the proper demand creation and presentation of the right time needs is also employed in these domains at the same time. Additionally, it is employed in

- financial forecasting,
- stock price estimation,
- management of new investments,
- portfolio determination,
- creation of marketing strategies,
- risk analysis,
- selection of the best human resources for businesses,
- estimation of credit and credit card fraud, credit limit determination, and fee management.

Applications for data mining are widely utilized to enhance the efficiency of various key business operations in the banking industry [12, 13]. Some banks employ data mining techniques to gather information on client behavior models and give suitable and effective promotions.

➢ *Telecommunication Sector*

Data mining can be used in the telecommunications industry to perform several functions of which may include:

- predict mobile user activity in the communications sector,
- determine future mobile user actions,
- detect frauds,
- minimize much of human-based analysis,
- identify the factors that cause customers to call more frequently at certain times;
- identify user templates for social network usage (19);
- identify potential clients using demographic data (20);
- identify the characteristics of new prospects.

➢ *Health sector*

For the health industry, data mining methods and application tools are more beneficial. Applications for data mining are widely employed to simplify the analysis of healthcare data transactions in the health industry [21–24]. Data mining is used in the healthcare industry to:

- diagnose diseases;
- determine appropriate treatments for diseases;
- estimate resource use and patient counts in hospitals;
- determine the efficacy of hospital-based treatments;
- classify patient data according to factors like age, gender, race, and treatment; identify high-risk factors for surgeries;
- prevent fraud in hospital spending.

➢ *Construction Sector*

Construction, project management, hydraulics, applications for occupational health and safety, analysis of earthquake data, groundwork investigations, and many more disciplines all employ data mining. In light of this research, it has been discovered that data mining techniques have been applied in the construction sector for the implementation of the following processes:

- create the information classification scheme in project documents [26],
- identify tax-related corruption,
- forecast the impact of changes in the tax system on the budget,
- identify waste and prevent damage caused by waste,
- estimate the cost of highway construction [27],
- estimate population, forecast the weather, and identify new job opportunities,
- To measure work productivity.

➢ *Public Sector*

In the public sector, data mining is frequently utilized to forecast issues with public safety and security. The public sector has several potential to make better judgments with the application 0f data mining tools. These choices are supported by broad trends deduced from historical data and experience [25]. In addition, it is used to:

- identify tax-related corruption,
- forecast how changes in the tax system will affect the budget,
- identify waste and prevent damage caused by waste,
- estimate population,
- forecast the weather,
- identify new job opportunities,
- measure employee performance and manage business processes,
- categorize public spending, plan the efficient use of resources, anticipate future public investment, examine data from the military sector, and identify criminals most likely to conduct crimes that endanger public safety.
- predict the future of public investment,
- analyze the data in defense of the future of public investment, measure worker productivity,
- manage business processes,
- classify public expenditures, plan the proper use of resources, define the characteristics of occupational accidents in the construction industry, and determine the compressive strength of concrete.

➢ *Engineering and Science Sector*

Numerous scientific disciplines, including astronomy, bioinformatics, computers, criminology, engineering, geosciences, mathematics, and software, have amassed vast amounts of data. Data mining has various advantages for the engineering and scientific fields, such as:

- controlling the software development life cycle for businesses,
- lowering the number of activities,
- accelerating software development,
- saving time and effort.
- by using anticipated analysis, firms may get a competitive edge,
- improve their industrial processes,
- analyze biological literature,
- use remote sensing, and analyze soil quality, among other things.
- classifying the sequences in bioinformatics, obtaining high-quality seed [20], evolving new crop breeds [20],
- classifying astronomical objects, and detecting patterns in crime.

These applications in the engineering and scientific fields support users by illuminating many aspects of current engineering software development processes and aiding in future decision-making processes.

III. CONCLUSION

In this paper, the techniques of data mining and application areas in some selected sectors have been reviewed. In various corporate fields, data mining is important for discovering patterns, forecasting, learning, etc. Classification, clustering and other data mining techniques and algorithms aid in identifying patterns that can be used to predict future business trends,

Data mining has wide application domain almost in every sector where the data is generated that is why data mining is regarded as one of the most significant frontiers in database and information systems and one of the most promising interdisciplinary developments in Information Technology.

REFERENCES

[1]. Abdullayeva, O., & Engalichev, M. (2022, October 26). Artificial intelligence systems. *Значение Цифровых Технологий В Изучении Истории Узбекистана*, *1*(01), 382–385. https://doi.org/10.47689/.v1i01.13612

[2]. Covid-19 vaccines and treatments: we must have raw data, now. (2022, January 24). *BMJ*, o189. https://doi.org/10.1136/bmj.o189

[3]. YAVUZ, Z. (2021, December 31). A Classification and Clustering Approach Using Data Mining Techniques in Analysing Gastrointestinal Tract. *International Scientific and Vocational Studies Journal*, *5*(2), 254–265. https://doi.org/10.47897/bilmes.1029171

[4]. Carlsen, L., & Bruggemann, R. (2022, July 22). Partial Order as Decision Support between Statistics and Multicriteria Decision Analyses. *Standards*, *2*(3), 306–328. https://doi.org/10.3390/standards2030022

[5]. Akpınar, H., 2014. DATA Veri Madenciliği Veri Analizi,Papatya Yayıncılık Eğitim, 1. Basım, İstanbul.

[6]. Ross, S. M., 2012. Introduction to Probability for Engineers and Scientists, Academic Press.

[7]. Kaya Keleş, M., 2016. İstatiksel Yöntemler ile Veri Madenciliği Yöntemlerinin Karşılaştırması (A Comparison of Statistical Methods and Data Mining Methods), Sosyal Bilimler Metinleri, 2016 Aralık ICOMEP Special Issue, pp. 20-24.

[8]. Özkan, Y., 2008. Veri Madenciliği Yöntemleri, Papatya Yayıncılık Eğitim, İstanbul.

[9]. Kleinberg, J., Papadimitriou, C., Raghavan, P., 1998. A microeconomic view of data mining, Data mining and knowledge discovery, 2.4, pp. 311-324.

[10]. Larose, D. T., 2005. Discovering Knowledge in Data, A John Wiley & Sons, Inc., Publication, New Jersey.

[11]. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P., 1996. From Data Mining to Knowledge Discovery in Databases, American Association for Artificial Intelligence, 17(3), pp. 37-54.

[12]. Sümersan Köktürk, M., Dirsehan, T., 2012. Veri Madenciliği ile Pazarlama Etkileşimi, Nobel Yayın, 1st Edition.

[13]. Kaya, M., Özel, S. A., 2014. Açık Kaynak Kodlu Veri Madenciliği Yazılımlarının Karşılaştırılması (Comparison of Open Source Data Mining Software), Akademik Bilişim'14 -

[14]. XVI. Akademik Bilişim Konferansı (AB2014), AB2014 Proceedings, 47-53.

[15]. Silahtaroğlu, G., 2008. Kavram ve Algoritmalarıyla Temel Veri Madenciliği, Papatya Yayıncılık Eğitim, İstanbul.

[16]. Pulakkazhy, S., Balan R.V.S., 2013. Data Mining In Banking And Its Applications-A Review, Journal of Computer Science, 9(10), pp. 1252-1259.

[17]. Bhambri, V., 2011. Application of Data Mining in Banking Sector, International Journal of Computer Science and Technology (IJCST), 2(2), pp. 199-202.

[18]. Kurt, Ç., Erdem, A., 2012. Öğrenci Başarısını etkileyen Faktörlerin Veri Madenciliği Yöntemleriyle İncelenmesi,Politeknik Dergisi, 15(2), pp. 111-116.

[19]. Ayık, Y. Z., Özdemir, A., Yavuz, U., 2007. Lise Türü ve Lise Mezuniyet Başarısının Kazanılan Fakülte ile İlişkisinin Veri Madenciliği Tekniği ile Analizi, Sosyal Bilimler EnstitüsüDergisi, 10(2), pp. 441-454.

[20]. Savaş, S., Topaloğlu, N., Yılmaz, M., 2012. Veri Madenciliği ve Türkiye'deki Uygulama Örnekleri, İstanbul Ticaret Üniversitesi Fen Bilimleri Dergisi, 21, pp. 1-23.

[21]. Çiftçi, S., 2006. Uzaktan Eğitimde Öğrencilerin Ders ÇalışmaEtkinliklerinin Log Verilerinin Analiz Edilerek İncelenmesi, Yüksek Lisans Tezi, Gazi Üniversitesi, Eğitim Bilimleri Enstitüsü.

[22]. Dolgun, M. O., Özdemir, T. G., Deliloğlu, S., 2007. Öğrenci Seçme Sınavında (ÖSS) Öğrencilerin Tercih Profillerinin Veri Madenciliği Yöntemleriyle Tespiti, Bilişim 07 Kongresi, Ankara.

[23]. Bozkır, A. S., Mazman, S. G., Sezer, E. A., 2010. Identification of User Patterns in Social Networks by Data Mining Techniques: Facebook Case, 2nd International Symposium on Information Management in a Changing World", Hacettepe University, Ankara, pp. 145-152.
[24]. Shahbaz, M., Rahman, M., 2008. Data mining for engineering sector in pakistan: Issues and implications, Proceedings of the World Congress on Engineering and Computer Science 2008 (WCECS 2008), San Francisco, USA, ISBN: 978-988-98671-0-2.
[25]. Avşar Aydın, E., Kaya Keleş, M., 2017. Breast cancer detection using K-nearest neighbors data mining method obtained from the bow-tie antenna dataset, International Journal of RF and Microwave Computer-Aided Engineering.Durairaj, M., Ranjani, V., 2013. Data Mining Applications In Healthcare Sector: A Study, International Journal Of Scientific & Technology Research, 2(10), pp. 29-35.
[26]. Avşar Aydın, E., 2014. Meme Kanseri Tespitinde Mikrodalgaların Önemi ve Kanserli/Sağlıklı Meme Dokularının Yapay Zeka Algoritmaları ile Tanımlanabilmesi, Çukurova Üniversitesi Mühendislik-Mimarlık Fakültesi Dergisi, 29(2), pp. 27-38.
[27]. Aydın, A., AVŞAR AYDIN, E., 2017. Evaluation of Limestone Layer's Effect for UWB Microwave Imaging of Breast Models Using Neural Network, Technical Journal, 11(1-2).
[28]. Wang, J, Hu, X., Zhu, D., 2008. Data Mining in Public Administration, Handbook of Research on Public Information Technology, IGI Global, Chapter LI, pp.556-567.

Author Profile

Grace Etiowo Jackson
She is a PhD student of University of Nigeria, Nsukka, Nigeria.
She has MSc degree (Information Technology) from National Open University of Nigeria (2016) . She is working as the Chief System Analyst, ICT Department, National Business and Technical examination's Board (NABTEB), Benin City, Nigeria.