International Journal of Innovative Research in Computer Science & Technology (IJIRCST)
ISSN: 2347-5552, Volume-10, Issue-2, March 2022
https://doi.org/10.55524/ijircst.2022.10.2.18
Article ID IRPV1039, Pages 92-95
www.ijircst.org

A study of Data Mining Techniques and Challenges

Dr. Suresh Kaswan,

Assistant Professor Department of Computer Science & Engineering, RIMT University, Mandi Gobindgarh, Punjab, India

Correspondence should be addressed to Dr. Suresh Kaswan; sureshkaswan@rimt.ac.in

ABSTRACT- In digital era, such as now, expansion of data in databases is quite quick; everything linked to technology, such as social media, financial technology, & scientific data, all contribute significantly to data growth. Because of enormous growth of information in this age of networking & info distribution, manually evaluating, categorising, & summarising data is difficult. As a result, subjects like big data & data mining are frequently explored. Data mining is procedure for dig out info from huge amounts of data in order to create a pattern or anomaly. In order to create innovative approaches for incorporating uncertainty management into data mining, this research looks into basics of data mining as well as existing research on integrating uncertainty into data mining. Management of indeterminate data, which might be instigated by obsolete resources, specimen mistakes, or inaccurate calculations, is among most difficult issues for technologies of data mining. Development of noval approaches for adding uncertainty supervision into data mining will be a focus of future study.

KEYWORDS- Algorithms, Classification, Clustering, Data Mining, Database.

I. INTRODUCTION

Advancement of information technology (IT) has resulted in numerous databases & massive volume of information in a wide range of sectors. Database & information systems research has yielded a technique for collecting & manipulating this important data for future decision-making. Process of obtaining useable trends & patterns from massive volumes of data is known as facts mining. This technique is referred to by several terms, including knowledge discovery, data mining, knowledge extraction, & data/pattern analysis.[1].

A. Background

Mother of invention is need. Our forefathers have looking aimed at meaningful info through data since dawn of time[2]. Nevertheless, with today's ever-increasing increased data, additional automatic & efficient mining approaches are requisite. In 1700s, Bayes' orem was employed to detect patterns from data, & in 1800s, regression analysis was applied. Because of quantity, universality, & ever-increasing power of digital technology, info accumulation surged substantially after 1900s[3]. As sets of data are growing in magnitude & complication, direct, hands-on analysis of data has progressively amplified by indirect, computational data acquisition. Other developments in computer science, like neural network models, segmentation, & evolutionary computation in 1950s, decision trees in 1960s, & support vector machines (SVM) in 1980s, have aided this[4].

Act of smearing these approaches to data with goal of finding unseen patterns is known as data mining [5]. Since many years, numerous industries, including businesses, researchers, & governments, have used data mining or data mining technology. It is used in sifting through enormous data sizes in order to develop market research reports, like air passenger trip statistics, demographic information, & market information, although this report is not referred to data mining[6].

It typically consist of 4 types of tasks:

- Classification, which divides info in predefined groups[7].
- Clustering, that is analogous to classification but does not divide info in predefined clusters[8]
- Regression, that tries finding best function to model data with least amount of error[9]
- Association rule learning, which looks for patterns in data[10].

It is logical technique for searching through vast quantities of info to uncover relevant info. objective of this way is discovering formerly undiscovered patterns[11]. Once these patterns have been discovered, they may be utilised to make specific decisions about how to grow their company. Exploration, Pattern Identification, & Deployment are three phases involved[12].

Exploration: Data exploration begins with cleaning & transformation of data into a new format, followed by identification of significant factors & type of data depending on problem[13].

Pattern Identification: second stage is to develop pattern identification after data has been investigated, distinguished, & specified for precise variables. Identify & choose patterns that provide most accurate predictions[14].

Deployment: Patterns are applied to get desired result[15].

B. Data Mining Algorithms & Techniques

Several algorithms & techniques involving Classification, Clustering, Regression, Artificial Intelligence, Neural Networks, Association Rules, Decision Trees, Genetic Algorithm, Nearest Neighbour method etc., are used for knowledge discovery from databases[16].

1) Classification

Furthermost common data mining strategy is classification, that employs group of pre-classified specimens for developing model capable of categorising entire populace of information[17]. This type of study

lends itself particularly well to fraudulent detection & credit risk concerns. Classification algorithms depending on decision trees or neural networks are frequently used in this strategy[18]. Data categorization process incorporates both learning & classification. In Learning, classification algorithm evaluates training data. Classification findings are used to determine correctness of classification rules. If precision is adequate, principles can apply to noval information tuples[19]. For misuse detection software, this would comprise detailed records of both duplicitous & legal activity analysed on record-by-record base. Classifier-training approach uses these pre-classified instance to define group of parameters requisite for proper discernment. method encodes these parameters.[20]. Categories of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM)
- Classification Based on Association

2) Clustering

It is process of identifying things that belong to comparable groups. Users may use clustering algorithms to discover entire distribution patterns & correlations between data attributes, as well as to detect dense & sparse locations in object space. Classification can be used to distinguish groups & classes of objects, but it takes time; thus, clustering could be used as a pre-processing tactic for attribute subdivision selection & classification. For instance, to identify genes with similar activities or to group customers based on their buying behaviour. Types of clustering methods[21]:

- Partitioning Methods
- Hierarchical Agglomerative (divisive) methods
- Density based methods
- Grid-based methods
- Model-based methods

3) Predication

Regression method may be used to forecast future. Regression analysis may use for modelling affiliation among 1 or more self-governing variable & reliant variables. Self-governing variables are characteristics that have previously been identified in data mining, whilst retort variables are whatever user wish to foretell. Unluckily, countless real-world challenges aren't amenable to straightforward forecasting. Sales volumes, product failure rates, & stock prices, for example, are challenging to forecast due to intricate interactions among many predictor factors. To anticipate future values, more complicated approaches might be required. When it comes to regression & classification, same model types are frequently utilised. CART (Classification & Regression Trees) decision tree technique, for example, may be use for create classification trees (to categorise categorical answer variables) as well as regression trees (to forecast continuous response variables). Both classification & regression models may be created using neural networks. Types of regression methods:

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression

4) Association rule

Purpose of affiliation & correlation is locate similar item set discoveries in huge data cluster. These kind of data assists businesses in deciding such as portfolio design, cross-marketing, & consumers' shopping activity recognition. Association rule algorithms must capable of producing rules with levels of confidence below one. Nevertheless, number of possible Association Rules for a large data is often rather large, & a significant proportion of them are of little (if any) utility. Categories of association rule:

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

5) Neural networks

It is made up of interconnected input/output units, each of which has own weight. In during process of learning, system changes weights to properly predict class labels of input tuples. Neural networks have an extraordinary ability of extracting patterns & identify trends from complicated or imprecise data, & these might used for extracting patterns & detecting trends that are just too complicated for humans or other computer technologies to find. These are ideal for inputs & outputs with a constant value. For instance, handwritten personality rearrangement, training machine to speak English text, & a variety of other real-world business challenges have all been effectively implemented in a variety of sectors. Neural networks excel in detecting patterns & trends in data, making them ideal for predicting & forecasting. Categories of neural networks

- Back Propagation

C. Data Mining Applications

It is relatively novel idea which is yet to grow & succeed. This is already being utilised on a daily basis by a range of sectors. Retail outlets, hospitals, banks, & insurance companies are just a few examples. Most of these companies are combining data mining with some other critical technologies such as statistics & pattern recognition. It might be used to identify patterns & correlations which would otherwise be hard to discover. Numerous businesses utilise this software since it permits them for learning additional about their clients & make better advertising decisions.

A human body contains around 100,000 genes, each of which is made up of hundreds of individual nucleotides organised in a certain order. There are unconstrained no. of traditions to arrange & sequence these nucleotides to generate different genes. Data mining may be used to examine sequential patterns, look for similarities, & find specific gene sequences linked to certain illnesses. Data mining technologies will become increasingly important in creation of novel medicines & cancer treatments in future.

Financial data gathered in banking & financial industries is frequently full, dependable, & of high quality, making

systematic data analysis & data mining easier. Customers may be classified & clustered for targeted marketing, money laundering & other financial crimes can be detected, & data warehouses can be designed & built for multidimensional data analysis.

Because it accumulates massive volumes of data on consumer purchasing history, consumption, & sales & service records, retail business is a key application area for data mining. Customer buying behaviours may be identified, purchase patterns can be discovered, & consumption trends can be predicted using data mining in retail industry. Data mining technology aids in development of efficient products transportation & distribution policies, as well as lower company costs.

It in telecommunications business may aid in understanding of company, identification of telecommunication trends, uncovering of fraudulent activities, better utilisation of resources, & improvement of service quality. Multidimensional telecommunication data analysis, fraudulent pattern analysis, & identification of odd patterns, along with multidimensional association & sequential pattern analysis, are examples of typical applications[22].

6) FBTO Dutch Insurance Company

i. Challenges

- For save money on direct mail.
- Make marketing efforts more efficient.
- Escalation of cross-selling to prevailing clients by utilising inbound channels such as company's sales centre & internet for a one-year evaluation of solution's efficacy.

ii. Results

- Assisting marketing team in predicting efficacy of their efforts.
- Made marketing campaign design, optimization, & implementation more efficient.
- Mailing expenses were reduced by 35%.

7) Conversion rates have increased by 40%. Standard Life Mutual Financial Services Companies

i. Challenges

- Recognise main characteristics of customers who are interested in hypothecation offer.
- Cross-selling of Standard Life Bank products to other Standard Life firms' customers.
- Create remortgage model that can used on group's website to assess cost-effectiveness of mortgage business Standard Life Bank accepts.

ii. Results

- Created propensity model for Standard Life Bank's mortgage offer, recognising important customer categories which can apply to whole prospect pool.
- Identified most important factors to consider when buying a reportage product.
- Achieved a nine-fold increase in response with model compared to control group.
- Received £33 million (about $47 million) in mortgage application fees.

8) Challenges of Data Mining

Researchers & developers face various needs & problems when it comes to efficient & effective data mining in huge datasets. Data mining technique, user engagement, performance & scalability, & processing of a wide range of data kinds are among challenges at h&. Or concerns include study of data mining applications & their societal consequences.

II. LITERATURE REVIEW

Han & Kamber discussed Data mining functions including data characterisation, data discernment, connotation examination, cataloguing, gathering, outlier analysis, & data evolution analysis. Summary of common physiognomies or attributes of target class of data is called data characterisation. General characteristics of target class objects are compared to general characteristics of items from one or more opposing classes in data discrimination. Identification of association guidelines indicating attribute-value circumstances that befall often together in particular collection of data is known as association analysis. Process of establishing collection of models or functions that define & distinguish data classes or ideas in order to use model to forecast class of objects whose class label is unknown is known as classification. Clustering examines data items without referring to a pre-defined class model. Outlier & data evolution analysis are methods for describing & modelling regularities or patterns in behaviour of things that vary over time[23].

Smyth P et al. addressed data mining principles, explaining how data mining & knowledge discovery in databases have recently gotten lot of interest from researchers, industry, & media. This page gives an overview of this new area, explaining how data mining & database knowledge discovery are connected to each other as well as to related fields like machine learning, statistics, & databases. article discusses specific real-world applications, data-mining approaches, difficulties in real-world knowledge discovery applications, as well as current & future research initiatives in subject[24].

Silwattananusarn T discussed Data Mining & Its Uses in Knowledge Administration in which author described how Data mining is a critical stage in information discovery process in databases, & it is a key subfield in knowledge management. In future decades, data mining research will continue to develop in business & learning organisations. Uses of data mining techniques that have been developed to help knowledge management process are explored in this review article. From 2007 to 2012, journal articles indexed in Science Direct Database were examined & categorised. Findings are discussed under four headings: knowledge resource; knowledge categories &/or datasets; data mining tasks; & data mining methodologies & applications in knowledge management. Definition of data mining & data mining functionality are briefly described in first section of article. Reasons for knowledge management & key knowledge management instruments used in knowledge management cycle are then discussed. Finally, use of data mining tools in knowledge management process is described & addressed[25].

III. DISCUSSION

This paper solely focuses on several aspects of data mining. It is logical technique for probing through vast quantities of data for uncovering relevant information. objective of this method is to discover previously undiscovered patterns. Once these patterns have been discovered, they may be utilised to make specific decisions about how to grow their company. Exploration, Pattern Identification, & Deployment are three phases involved. These phases along with many other concepts have been discussed in this paper. This paper discusses several algorithms & techniques of data mining. It discusses background of data mining. It discusses several real life applications of data mining along with several challenges faced in data mining.

IV. CONCLUSION

This article provides an overview of data mining, which is procedure of retrieving beneficial information from vast volumes of data stored in databases. It also covers basics of data mining as well as approaches for incorporating uncertainty in data mining, such as K-means algorithm. It is demonstrated that data mining technology may be functional to extensive range of real-world applications, including biomedical & DNA data analysis, financial data analysis, retail business, & telecommunications. Management of uncertain data, which might be caused obsolete resources, sampling mistakes, or inaccurate calculations, is one of most difficult issues for data mining technologies. Development of novel approaches for adding uncertainty management into data mining will be a focus of future study.

REFERENCES

[1] Y. N & M. S, "A Review on Text Mining in Data Mining," Int. J. Soft Comput., 2016, doi: 10.5121/ijsc.2016.7301.

[2] S. Gupta & G. Khan, "MHCDA: A proposal for data collection in Wireless Sensor Network," 2017, doi: 10.1109/SYSMART.2016.7894517.

[3] D. Sehgal & A. K. Agarwal, "Real-time sentiment analysis of big data applications using twitter data with Hadoop framework," 2018, doi: 10.1007/978-981-10-5699-4_72.

[4] S. VijayGaikwad, A. Chaugule, & P. Patil, "Text Mining Methods & Techniques," Int. J. Comput. Appl., 2014, doi: 10.5120/14937-3507.

[5] J. Apostolakis, "An introduction to data mining," Struct. Bond., 2010, doi: 10.1007/430_2009_1.

[6] G. Khan, K. K. Gola, & M. Dhingra, "Efficient techniques for data aggregation in underwater sensor networks," J. Electr. Syst., 2020.

[7] M. M. Gupta, S. Jankie, S. S. Pancholi, D. Talukdar, P. K. Sahu, & B. Sa, "Asynchronous environment assessment: A pertinent option for medical & allied health profession education during the covid-19 p&emic," Education Sciences. 2020, doi: 10.3390/educsci10120352.

[8] M. H. F. Siddiqui & R. Kumar, "Interpreting the Nature of Rainfall with AI & Big Data Models," 2020, doi: 10.1109/ICIEM48762.2020.9160322.

[9] S. Goel, R. K. Dwivedi, & A. Sharma, "Analysis of social network using data mining techniques," 2020, doi: 10.1109/SMART50582.2020.9337153.

[10] D. Gupta et al., "Musculoskeletal pain management among dentists: An alternative approach," Holist. Nurs. Pract., 2015, doi: 10.1097/HNP.0000000000000074.

[11] P. Gupta & N. Tyagi, "An approach towards big data - A review," 2015, doi: 10.1109/CCAA.2015.7148356.

[12] K. S. Deepashri & A. Kamath, "Survey on Techniques of Data Mining & its Applications," Int. J. Emerg. Res. Manag. Technol., 2017.

[13] M. S&hu, Jayan&, B. Rawat, & R. Dixit, "Biologically important databases available in public domain with focus on rice," Biomedicine (India). 2017.

[14] G. Mariscal, Ó. Marbán, & C. Fernández, "A survey of data mining & knowledge discovery process models & methodologies," Knowledge Engineering Review. 2010, doi: 10.1017/S0269888910000032.

[15] S. Kumar, J. Shekhar, & J. P. Singh, "Data security & encryption technique for cloud storage," 2018, doi: 10.1007/978-981-10-8536-9_19.

[16] M. S. Solanki, D. K. P. Sharma, L. Goswami, R. Sikka, & V. An&, "Automatic Identification of Temples in Digital Images through Scale Invariant Feature Transform," 2020, doi: 10.1109/ICCSEA49143.2020.9132897.

[17] L. Goswami, M. K. Kaushik, R. Sikka, V. An&, K. Prasad Sharma, & M. Singh Solanki, "IOT Based Fault Detection of Underground Cables through Node MCU Module," 2020, doi: 10.1109/ICCSEA49143.2020.9132893.

[18] K. Sharma & L. Goswami, "RFID based Smart Railway Pantograph Control in a Different Phase of Power Line," 2020, doi: 10.1109/ICIRCA48905.2020.9183202.

[19] M. Khatri & A. Kumar, "Stability Inspection of Isolated Hydro Power Plant with Cuttlefish Algorithm," 2020, doi: 10.1109/DASA51403.2020.9317242.

[20] P. Guleria & M. Sood, "Data Mining in Education : A Review on the Knowledge Discovery Perspective," Int. J. Data Min. Knowl. Manag. Process, 2014, doi: 10.5121/ijdkp.2014.4504.

[21] W. Ghai, S. Kumar, & V. A. Athavale, "Using gaussian mixtures on triphone acoustic modelling-based punjabi continuous speech recognition," 2021, doi: 10.1007/978-981-15-1275-9_32.

[22] F. Xiao & C. Fan, "Data mining in building automation system for improving building operational performance," Energy Build., 2014, doi: 10.1016/j.enbuild.2014.02.005.

[23] J. Kaur & N. Madan, "Association Rule Mining: A Survey," Int. J. Hybrid Inf. Technol., 2015, doi: 10.14257/ijhit.2015.8.7.22.

[24] U. Fayyad, G. Piatetsky-Shapiro, & P. Smyth, "From data mining to knowledge discovery in databases," AI Mag., 1996.

[25] T. Silwattananusarn, "Data Mining & Its Applications for Knowledge Management : A Literature Review from 2007 to 2012," Int. J. Data Min. Knowl. Manag. Process, 2012, doi: 10.5121/ijdkp.2012.2502.