Review On Challenges and Issues in Data Mining

Omkar K. Khadke[1], Lowlesh N. Yadav[2], Neehal B. Jiwane[3]

B-Tech Final Year Student, Computer Science And Engineering, Shri Sai College Of Engineering And Technology, Bhadrawati, Maharashtra, India[1]

Assistant Professor, Computer Science And Engineering, Shri Sai College Of Engineering And Technology, Bhadrawati, Maharashtra, India[2]

Assistant Professor, Computer Science And Engineering, Shri Sai College Of Engineering And Technology, Bhadrawati, Maharashtra, India[3]

Abstract: Data Mining data and data from high databases has been recognized by several researchers as a key analysis topic in machine learning and info system and in several industrial corporations as a crucial space with a chance of major revenues .we gift a two-dimensional read of information mining. the foremost dimensions area unit data, knowledge, technologies and application. Researchers in some totally different fields have shown their nice interest in data mining. during this section, we tend to in brief define methodology, user interaction. data processing analysis has powerfully impact society and can still do thus within the future.

Keywords: Data Mining, Machine Learning, Data Mining Issues.

1. INTRODUCTION

Data mining is that the method of motion numerous queries and extracting helpful info, patterns, and trends typically antecedently unknown from massive quantities of knowledge probably keep in databases. basically, for several organizations, the goals of knowledge mining embrace up selling capabilities, detection abnormal patterns, and predicting the long run supported past experiences and current trends. there's clearly a requirement for this technology. There square measure massive amounts of current and historical information being keep. Therefore, as databases become larger, it becomes progressively tough to support higher cognitive process. additionally, the information might be from multiple sources and multiple domains. there's clear have to be compelled to associatealyze the information to support designing and alternative functions of an enterprise. numerous terms are wont to talk to data processing. These embrace knowledge/data/information discovery and knowledge/data/information extraction. so data processing is that the natural evolution of question and news tools. Everyone, World Health Organization creates queries and reports, advantages from having data processing capabilities

2. CHARACTERISTIC OF DATA MINING

Characteristics of information Mining:
Data mining service is a simple sort of military operation methodology whereby that all the relevant data goes through some style of identification method. And eventually at the tip of this method, one will confirm all the characteristics of the information mining method.

2.1. Enhanced quantities of data: In earlier days, the information mining system will be determined with the assistance of their shoppers and customers, however in today's date, one will acquire any range of data while not the assistance of these shoppers. Moreover, once this sort of revolution within the mining system, it conjointly extra a new downside which is massive quantities of labor. With the assistance of this data technology, one will acquire an oversized range of data with none additional burden or bother.

2.2. Provides incomplete data: Most of the folks give incomplete data concerning themselves in a number of the survey conducted with the assistance of information mining systems. Therefore, folks ignore the worth of their data which is why they supply incomplete data concerning themselves in those surveys conducted for the good thing about the mining

systems. Moreover, these mining systems modified the attitude of individuals and since of that, folks worry the exchange of their personal data.

2.3. difficult information structure: Data mining could be a kind whereby that all the knowledge is gathered and incorporated with the assistance of data assortment techniques. These data collection techniques area unit additional of manual and rest area unit technological. Therefore, most of the understanding and determination of those mining will be a touch difficult than different structures of data technology.

3. DATA MINING APPLICATIONS

Data mining is generally employed by several of the large gaints within the info technology sector and conjointly some little industries by creating use of their own techniques. a number of the favored domains ar, Market Analysis and Management Corporate Analysis & Risk Management Fraud Detection

3.1. market research and Management: The following mentioned ar the varied fields of the market wherever the information mining method is effectively used,

- Customer identification
- Finding client necessities
- Cross-market analysis
- Target selling
- Determining client buying pattern
- Provides outline info

3.2. company Analysis and Risk Management: The following mentioned ar the varied fields of the company sector wherever the information mining method is effectively used,

- Finance designing
- Asset analysis
- Resource designing
- Competition

3.3. Fraud Detection: Frauds and malware is one in every of the foremost dangerous threats on the net. it's virtually a sort of crime that's increasing day once day. The fraud detection method is chiefly used through MasterCard services and telecommunication. With the assistance of the services most of the necessary info like length of the decision, location, the time and day etc is non inheritable that helps in success.

4. DATA MINING CHALLENGES

These days data processing and knowledge revealing area unit developing crucial innovations for researchers and businesses in varied areas. data processing was forming into a setup and confided up to the mark, heretofore forthcoming data processing challenges should be tackled.

4.1. Security and Social Challenges
Dynamic techniques area unit done through knowledge assortment sharing, which needs spectacular security. non-public data concerning individuals and touchy data is gathered for the client's profiles, shopper normal of conduct understanding—illicit admittance to data and therefore the secret plan of knowledge turning into a big issue.

4.2. Clanging and Incomplete knowledge
Data Mining could be a thanks to get data from Brobdingnagian volumes of information. This gift reality of knowledge is clanging, incomplete, and heterogeneous. knowledge in Brobdingnagian amounts often are unreliable or inaccurate. These problems may be owing to human mistakes, blunders, or errors within the instruments that live the info.

4.3. Distributed knowledge
True knowledge is generally place away at varied stages in distributed process conditions. it's going to get on the web, individual systems, or maybe databases. it's basically arduous to hold all {the knowledge|the info|the information} to a unified data archive chiefly owing to technical and structure reasons.

4.4. Complicated knowledge

True knowledge is heterogeneous, and it's going to be media knowledge, as well as tongue text, statistic, spatial knowledge, temporal knowledge, complicated knowledge, audio or video, images, etc. it's actually arduous to alter these varied forms of knowledge and think about the required data. a lot of typically than not, new apparatuses and systems would want to be created to separate vital data.

4.5. Performance

The presentation of the info mining framework primarily depends upon the productivity of techniques and algorithms used. On the off likelihood that the techniques and algorithms planned aren't sufficient , at that time, it'll influence the presentation of the info mining live unfavorably.

4.6. Quantifiability and potency of the Algorithms

The Data Mining rule ought to be climbable and economical to disengage data from tremendous measures {of knowledge|of knowledge|of information} within the data set.

4.7. Improvement of Mining Algorithms

Factors, for instance, the issue of information mining approaches, the large size of the information, and therefore the entire knowledge flow, inspire the distribution and creation of parallel data processing algorithms.

4.8. Incorporation of information

In the event that information will be consolidated, a lot of correct and reliable data processing arrangements will be found. prognosticative tasks will create a lot of correct predictions, whereas descriptive tasks will come back up with a lot of helpful findings. Be that because it could, gathering and as well as foundation information is unpredictable.

4.9. Knowledge visualization

Data visualization could be a very important cycle in data processing since it's the foremost interaction that shows the output in an exceedingly respectable thanks to the shopper. the knowledge freed have to be compelled to pass away the importance of what it plans to pass away. However, ordinarily, it's actually arduous to handle the knowledge exactly and squarely to the top user. The output data and computer file being terribly effective, successful, and sophisticated knowledge perception strategies ought to be applied to create it fruitful.

4.10. Knowledge Privacy and Security

Data mining generally prompts important governance, privacy, and knowledge security problems. as an example, once a distributor investigates the acquisition details, it uncovers data concerning buying propensities and decisions of shoppers while not their authorization.

4.11. Interface

The information is set utilizing data processing devices is effective simply within the event that it's fascinating or a lot of all affordable by the shopper. From nice illustration translation of information, mining results will be expedited, and betters comprehend their conditions. several explorations area unit in deep trouble monumental knowledge sets that manipulate and show strip-mined information to urge an excellent perception.

4.12. Mining keen about Level of Abstraction

Data Mining measures ought to be community-oriented in light-weight of the very fact that it permits shoppers to specialize in example optimizing, presenting, and pattern finding for data processing keen about conveyance results back.

4.13. Integration of information

Previous data can be accustomed communicate examples to precise discovered patterns and direct the exploration method.

4.14. Mining Methodology Challenges

These difficulties area unit known with data processing strategies and their limits. Mining strategies that cause the difficulty area unit the management and handling of noise in knowledge, the spatiality of the domain, the range of information accessible, the flexibility of the mining methodology, and so on.

5. DATA MINING ISSUES

5.1. Mining Strategies & User Interaction Problems

The issues may be with finding out the best-fit methodologies of knowledge mining, that area unit association, classification, clump analysis, prediction, sequent patterns or pattern pursuit, call trees, outlier analysis or anomaly

detection, and neural network.The professionals with active expertise during this domain usually struggle with these problems whereas victimization the same data processing strategies.

5.1.1.Hard to Derive information for numerous Domains
The beneficiary may be associated with totally different industries and domains. thus do dissent their demand and information discovery, that need niche-based knowledge extraction, transformation, & loading of numerous forms like visuals, text, or numbers. This method covers a broad vary of information discovery processes, that may be a challenge.

5.1.2.Lack of Interactive Models
Interactive modelling makes pattern searches easier. For this purpose, datasets area unit extracted, refined, converted, and cleaned to make sure that they manufacture the intelligence that's needed through the mining request.

5.1.3.Lacking skilled information
For every level, from net knowledge extraction to process, & modelling, you would like knowledgeable matter specialists or knowledge miners. Their active expertise guides them through the complete discovery method, that is to work out patterns. Their qualification and operating expertise guarantee golf stroke patterns during a crisp and comprehensive format. Finding such professionals with answer ability is so Associate in Nursing uphill battle.

5.1.4.Ad hoc data processing isn't straightforward
Ad-hoc mining answers to a selected business demand or question, that insights will create it easier to abstract. The Structured search language is that the search language that supports the accidental information discovery method. This language guides users through mining tasks by mining or knowledge analysis corporations, that work on the optimized & versatile structure of records. This task is once more not very easy.

5.1.5.Comprehensiveness may be a Challenge
The visual presentation ought to be impactful to let the analyst simply see and extract insights in no time. Here again, the team ought to be competent to run totally different tools like Tableau, Sisence, Excel, etc. for a good visual presentation victimization charts, graphs, etc. Simply put, the presentation of models ought to be perceivable, that isn't sort of a walkover.

5.1.6.Noisy knowledge is tough to Handle
Noisy knowledge talk over with redundancies or useless knowledge within the info or massive knowledge. addressing the noises like duplicity, incomplete info, and errors may be an enormous challenge. But, going ahead with them will disturb the effectiveness of consequence. Result? The pattern are poor and useless.

5.1.7.Measuring Patterns
The patterns filtered, tested, & evaluated ought to be attention-grabbing as a result of they represent either intelligence or possible solutions or lack novelty.

5.2. Performance problems

It is necessary that the modelling ought to be versatile and qualified through quality tests. These models area unit seemingly to be AI via machine learning. Here, the performance problems should be acknowledged to sail across them with previous solutions.

5.2.1.Inefficiency and issue in measurability
Certainly, a info is needed, principally colossal-sized niche-based records. net knowledge extraction & capturing create it approach easier. although it's paper-bound, the OCR conversion and cleansing practices facilitate in making ready a info. But, there area unit challenges like king protea lure, captcha, and privacy settings which will hamper the provision of significant details. Sometimes, lacking tools may prove an enormous barrier to shaping the potency and measurability of knowledge mining.

5.2.2.Inability to figure with Parallel, Distributed, and progressive Algorithms
The metrics like prodigious size of databases, wide distribution, and complexness of knowledge mining strategies push to derive parallel and distributed data processing algorithms. These algorithms split records into elements, that area unit any processed within the same manner. Finally, the results of all elements area unit compiled along. this is often however the progressive algorithms still update databases while not mining the info once more from scratch.

5.3. Numerous knowledge varieties problems

Data has several faces. you'll notice it in its visual, audio, text, and numeric forms. process these differing types and so, mining could also be troublesome.

5.3.1.Dealing with relative and complicated sorts of knowledge

Your supply knowledge might have PDF, transmission objects, abstraction knowledge, temporal or alternative sorts of datasets. this is often so a bone-breaking expertise to form such a customary tool which will ideally method every kind of knowledge within the same approach.

So, you would like to customize or access a specifically designed tool to mine from a specific kind of datasets, that is a rich deal. And also, you must have the power to manage that tool.

5.3.2.Modelling from Heterogeneous Databases

As there area unit variety of sources to access knowledge like local area network or WAN, you'll not expect to own records in Associate in Nursing ideally similar type and format. it's just because of their storage during a structured, semi-structured, or unstructured type. Therefore, mining information from them isn't straightforward.

5.4. knowledge Security & Privacy

Personally, place able knowledge is sensitive and other people don't wish to share it with anyone. Here, a threat to its privacy & security may be a reason to significantly suppose.

5.4.1.Security

Mostly, knowledge area unit shared over the web, the cloud, and servers to make sure their access 24X7 remotely. This access may be dangerous if it's done through a public network, that isn't secure. Vulnerability poses an enormous risk. So, the inter changeableness of any record ought to be outlined through encoding.

5.4.2.Privacy considerations

Dynamic techniques area unit adopted to gather info from numerous resources, particularly from knowledge subjects. This assortment isn't unhazardous, as they carry in person place-able info. Hackers tend to interrupt in and exclude these credentials. Here, privacy controls, authorization, and knowledge compliance like GDPR seem during a major safeguarding role.

6. ADVANTAGES OF DATA MINING TECHNIQUES

There square measure many styles of advantages and blessings of information mining systems. one among the essential matters of those mining creates an entire structure of research of mining techniques.

6.1.It's useful to predict future trends: Most of the operating nature of the info mining systems carries on all the informational factors of the weather and their structure.One of the common advantages that may be derived with these data processing systems is that they will be useful whereas predicting future trends. which is sort of potential with the assistance of technology and behavioral changes adopted by the individuals.

6.2. It signifies client habits: For example, whereas operating within the selling trade one will perceive all the matters of client behaviour and their habits. which is feasible with the assistance of information mining systems. As these data processing systems handle all the knowledge deed techniques. it's useful keep track of client habits and their behavior.

6.3. Helps in call making: There square measure some folks that create use of those data processing techniques to assist them with some reasonably deciding. Nowadays, all the knowledge concerning something may be determined simply with the assistance of technology and equally, with the assistance of such technology one will create an explicit call concerning one thing unknown and surprising.

6.4. Increase company revenue: As it has been explained earlier that methoding} could be a process whereby that it involves some form of technology to accumulate some info concerning something potential. And this kind of technology makes things easier for his or her profit earning magnitude relation. As individuals will collect info concerning the marketed merchandise on-line, that eventually reduces the price of the merchandise and their services.

6.5. It depends upon market-based analysis: Data mining method could be a system whereby that all {the info|the knowledge|the data} has been gathered on the idea of market information. Nowadays, technology plays a vital role in everything which casualty may be seen in these data processing systems. Therefore, all the knowledge collected through these data processing is largely from selling analysis.

6.6. Fast fraud detection: Most components of {the data|the info|the info} mining method is largely from information gathered with the assistance of promoting analysis. With the assistance of such selling analysis, one also can establish those dishonest acts and merchandise obtainable within the market. Moreover, with the assistance of it one will perceive the importance of correct info.

7. DISADVANTAGES OF DATA MINING TECHNIQUES

Data mining technology are some things that helps one person in their {decision making|deciding|higher cognitive method} which deciding could be a process whereby that all the factors of mining is concerned exactly. And whereas the involvement of those mining systems, one will stumble upon many disadvantages of information mining and that they square measure as follows.

7.1. It violates user privacy: It is a far-famed incontrovertible fact that data processing collects info concerning individuals mistreatment some market-based techniques and data technology. And these methoding} process involves many numbers of things. But whereas involving those factors, data processing system violates the privacy of its user which is why it lacks within the matters of safety and security of its users. Eventually, it creates miscommunication between individuals.

7.2. Further extraneous information: The main functions of {the data|the info|the info} mining systems produce a relevant house for useful information. But the most drawback with these info collections is that there's a prospect that the gathering of data processes may be a touch overwhelming for all. Therefore, it's noticeably essential to take care of a minimum level of limit for all the info mining techniques.

7.3. Misuse of information: As it has been explained earlier that within the data processing system the chance of safety and security live square measure extremely minimal . which is why some will misuse this info to hurt others in their own manner. Therefore, the info mining system has to amendment its course of operating in order that it will scale back the magnitude relation of misuse of data through the mining method.

7.4. Accuracy of data: Most of the time whereas collection info concerning sure components one wont to ask for facilitate from their shoppers, however today everything has modified. And currently the method of data assortment created things simple with the mining technology and their ways. One of the foremost potential limitations of this data processing system is that it will offer accuracy of information with its own limits.

8.CONCLUSION

Data mining seeks to extract hidden information from large amount of information. methoding} is that the process of extracting and valuable attention-grabbing patterns from raw collection of information. data processing may be accustomed uncover patterns within the information however it's typically dispensed solely on the samples of information. This mining method are ineffective if the samples don't seem to be a decent illustration of the larger body of the info.And beside this, today's competition is one among the most necessary challenges facing by all organizations
and industries in data processing problems. that's onerous to seek out in a particular organization or business that has no rival to him.This paper describes numerous tasks; goals and limitations of information mining. Additionally-this paper additionally discussed regarding the varied valuable problems; future challenges and problems in field of information mining that I important to try to to more more practical analysis during this emerging field.

REFERENCES

[1] M. S. Chen, J. Han, and P. S. Yu. Data mining: An overview from a database perspective. IEEE Trans. Knowledge and Data Engineering, 8:866-883, 1996.

[2] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy. Advances in Knowledge Discovery and Data Mining. AAAI/MIT Press, 1996.

[3] Chieh-Yuan, T. and T. Min-Hong. A dynamic Web service based data mining process system. in Computer and Information Technology, 2005. CIT 2005. The Fifth International Conference on. 2005.

[4] Ding, P. A formal framework for Data Mining process model. In Computational Intelligence and Industrial Applications, 2009. PACIIA 2009. Asia-Pacific Conference on. 2009.

[5] Cooley, B. Mobasher and J. Srivsatava. Web Mining: Information and Pattern Discoveryon the Word Wide Web. In Proceedings of the 9th IEEE International Conference on Tools with AI (ICTAI,97), Nov. 1997.

[6] Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. SIGMOD'93.

[7] J. Bayardo. Efficiently mining long patterns from databases. SIGMOD'98.

[8] Huhtala, J. Kärkkäinen, P. Porkka, H. Toivonen. Efficient Discovery of Functional and Approximate Dependencies Using Partitions. ICDE'98.

[9] W. Shavlik and T. G. Dietterich. Readings in Machine Learning.Morgan Kaufmann, 1990.

[10] P.Tan, M. Steinbach, and V. Kumar. Introduction to Data Mining.Addison Wesley, 2005.

[11] S. M. Weiss and C. A. Kulikowski. Computer Systems that Learn: Classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, and Expert Systems.Morgan Kaufman, 1991.

[12] Xinjian Guo, Yilong Yin1, Cailing Dong, Gongping Yang, Guangtong Zhou,"On the Class Imbalance Problem" Fourth International Conference on Natural Computation, 2008.
[13]
[14] David P. Williams, Member, Vincent Myers, and Miranda Schatten Silvious, "Mine Classification With Imbalanced Data", IEEE Geosciences And Remote Sensing Letters, Vol. 6, No.3, July 2009.

[15] Chris Seiffert, Taghi M. Khoshgoftaar, Jason Van Hulse, Amri Napolitano "A Comparative Study of Data Sampling and Cost Sensitive Learning" , IEEE International Conference on Data Mining Workshops.15-19Dec.2008.
[16]
[17] Clifton, Christopher "Encyclopedia Britannica: Definition of Data Mining", Retrieved 2010-12-09.

[18] Ian H. Witten; Eibe Frank; Mark A. Hall, ―Data Mining: Practical Machine Learning Tools and Techniques (3rd Ed.)‖, Elsevier, 30 January 2011.

[19] Kantardzic, Mehmed ―Data Mining: Concepts, Models, Methods, and Algorithms‖, John Wiley & Sons, 2003.

[20] Günnemann, S.; Kremer, H.; Seidl, T. "An extension of the PMML standard to subspace clustering models". Proceedings of the 2011 workshop on Predictive markup language modeling - PMML '11.

[21] Ralf Mikut; Markus Reischl, "Data Mining Tools". Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery. Retrieved October 21, 2011.

[22] Domenico Talia; Paolo Trunfio"How distributed data mining tasks can thrive as knowledge services". Communications of the ACM , 2010 .

[23] Karl Rexer, Heather Allen, & Paul Gearan ―Data Miner Survey – Overcoming Data Mining's Top Challenges‖, 2011.

[24] Hans-Peter Kriegel , Karsten M. Borgwardt et.al ―Future trends in data mining‖, sprinzer, 2007.

[25] QIANG YANG,XINDONG WU, ―10 Challenging Problems in Data Mining Research‖.
[26] Michael Goebel,Le Gruenwald ―A Survey of Data Mining and Knowledge Discovery Software Tools‖, ACM, 1999.
[27] DianshengGuo, Jeremy Mennis, ―Spatial data mining and geographic knowledge discovery—An introduction‖, 2009 Elsevier
[28] Pavel Berkhin Accrue Software, Inc.‖Survey of Clustering Data Mining Techniques‖.