IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-6, Issue-8, Aug 2018 E-ISSN: 2347-2693

Survey on Data Mining Technique

Harmeet Kaur[1*], Jasleen Kaur[2]

[1*]M.tech Scholar, Dep. of CSE, Khalsa College of Engineering And Technology, Amritsar, India
[2]A.P, Dep. of CSE, Khalsa College of Engineering and Technology, Amritsar, India

[*]*Corresponding Author:er.harmeetsodhi430@gmail.com*

Available online at: www.ijcseonline.org

Accepted: 15/Aug/2018, Published: 31/Aug/2018

***Abstract*—**Now a Day, Many Companies and organizations are to make a large volume of information. In the enterprise, choice producers access from all sources and types of collection methods. The information warehouse is used for the enterprise for enhancing the selection-making. In aggressive commercial enterprise international, the values of strategic statistics techniques along with these are actually identified. The enterprise surroundings, the pace isn't the simplest key to competitiveness. To investigate the records, it needs the unique equipment are called facts mining things. This paper survey of the facts mining set of rules which include Clustering, Time series, Logistic Regression, Naïve Bayes and its programs within the exclusive areas.

***Keywords*—**Data mining, Clustering, Time series, Logistic Regression, Naïve Bayes

I. INTRODUCTION

The facts mining custom to assess the data. These data records may be used to grow the income and expenses. Then the customers to evaluate the records from extraordinary angles, the organization it, and encapsulate the members of the family diagnosed. Technically the statistics mining technique to locate the relationships between the numbers of fields in RDBMS. Data mining programming is one of the various logical instruments for dissecting information. data Mining is characterized as separating data from immense arrangements of information. At the end of the day, we can state that information mining is the method of mining learning from information. Rundown of steps engaged with the information revelation process –

- **Data Cleaning:** In this progression, the, and conflicting information are evacuated.
- **Data Integration:** In this progression, various information sources are consolidated.
- **Data Selection:** In this progression, information important to the examination assignment is recovered from the database.
- **Data Transformation:** In this progression, information is changed or merged into frames fitting for mining by performing outline or total activities.
- **Data Mining:** In this progression, canny techniques are connected with a specific end goal to separate information designs.
- **Pattern Evaluation:** In this progression, information designs are assessed.
- **Knowledge Presentation:** In this progression, learning is assessed too.

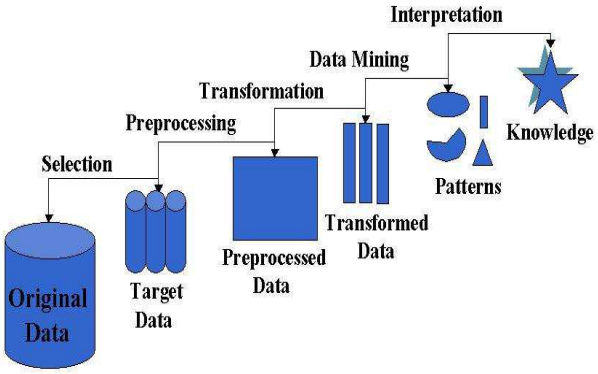

Figure 1. Example of a figure caption. *(figure caption)*

A. Data Mining Life Cycle

The records of Data Mining life cycle consists of six ranges. The order of the degrees is not rigid. Moving backward and forward among distinct levels is required. It relies upon at the result of each stage. The principal ranges are:

1. **Commercial enterprise expertise:** The records mining efforts the challenge requests and objects from a commercial enterprise, changing this expertise right into a statistics mining trouble definition. It's a far number one plan designed to obtain the goals.

2. **Facts knowledge:** It is an original statistics series, to accumulate the vital facts, to identify the information cost difficulties, to locate the inspiring subgroups to shape guesses for hidden information.

3. **Records preparation:** The information collects all records and makes the one of a kind movements constructing at the initial rare information.

4. Modeling: In the modeling stage, distinct modeling techniques are nominated. Its miles practical and their elements are standardized for quality ideas.

5. Assessment: The evaluation version is systematically valued and revised. Then, to execute the hypothesis model. To be precise it appropriately attains the industrial gadgets. At the last of this degree, an outcome to be reached at the statistics mining consequences.

6. Deployment: Once an assessment is widespread via the management, the ultimate step is the deployment step. To boom statistics, the information accelerated to be deliberate and present in a way that the consumer can use it. In these level may be making a documentary.

B. Data Mining Application

It is used for an extensive range of software. An information Data Mining application is followed via organizations. Those corporations are related groups, not disjoint corporations.

1. **Service providers:** Carrier carriers used for the mobiles, groups, and industries. It is used for records Mining to are expecting ‘churn', the technique used for whilst a patron left their agency to get their telephone or landline broadband from a new provider. They incorrect billing records, offerings interactions, website visits and online purchasing, they could spot clients who they concept had been in all likelihood to be expecting and start concentrating on promotions for nappies (diapers), cotton wool and so on.

2. **Crime Prevention:** Data mining application utilized in crime. Crime prevention corporations use analytics and statistics mining to spot tendencies throughout myriads of facts – supporting with the whole lot from where to set up police manpower. To search at a border crossing based totally on age/kind of car, variety/age of occupants, border crossing history. To take extremely in counters - terrorism sports. Then to present each purchase a probability score, the goal gives and incentives to customers they understand to be a higher chance of churning.

3. **Retail:** The store's segment clients into ‘Regency, Frequency, economic' (RFM) businesses. To goal advertising and promotions of the special corporations. A patron who spends less however frequently and last did so lately may be dealt with differently to a purchaser who spent long, however, handiest once, and additionally sometime in the past. The previous may additionally receive a loyalty, upsell and go sell gives, the latter may be supplied a win back deal.

4. **E-commerce:** Records mining software utilized in e-commerce. Many E-commerce agencies use statistics mining to offer pass-sells and upsell through their websites. One of the maximum well-known of these is, of a route, Amazon, who use sophisticated mining strategies to pressure there, ‘people who viewed that product, also preferred this' functionality.

5. **Supermarkets:** Data mining business software used in Supermarkets. Famously, grocery store loyalty card programmers are usually pushed generally, if no longer completely, by using the desire to collect comprehensive data about customers to be used in information mining.

As a last paragraph of the introduction should provide organization of the paper is organized as follows, Section I contains the introduction of Data Mining, Section II contain the related review, Section III contain the some techniques related to data mining, Section IV contain Gaps in the preview work, section V contain conclusion of research work with future directions.

II. RELATED WORK

Rajwinder Kaur et al. (Jan 2017) provided the thought of Classification. Classification is an important data mining technique centered on machine learning which is used to classify the each item on the bases of options that come with the item with respect to the predefined group of classes or groups and summarized various techniques which can be implemented for the classification such as for instance k-NN, Decision Tree, Naïve Bayes, SVM, ANN and RF. The techniques are analyzed and compared on the cornerstone of their advantages and disadvantages. Surekha Samsani et al. (Jan 2016) simulated the collective intelligence of honey bee swarms in nature called the Artificial Bee Colony optimization technique; for finding the minimal subset of features for the classification of Thyroid disease. The results demonstrated that the ABC algorithm for feature subset selection has generated an optimal set of features with the significant increase in the accuracy of the classification of Thyroid disease dataset. Mayuri Takore et al. (Mar 2016) concerned with using statistical and data mining methods to boost data evaluation on big data sets. Disease examination is one of many purposes wherever data mining methods are proving successful benefits and planned to understand the center diseases through data mining, Help Vector Unit (SVM), Genetic Algorithm, rough set idea, association rules and Neural Networks. Memorie Mwanza et al. (Mar 2016)

based on two areas. We were keeping (1) the baseline study that helped to ascertain the degree of the problems in scam detection for the taxpayers and (2) the automation and progress of the scam detection tool utilizing the benefits from the baseline study and generated by our tool showed increased precision and requires less time in order to find under and overpayments as outliers when comparing to the older methods. Sheenal Patel et al. (Mar 2016) Data mining has an infinite potential to utilize healthcare data more efficiently and effectually to predict a different kind of disease and featured various Data Mining techniques such as classification, clustering, association and also highlighted related work to analyzed and predict human disease. Siddharth S. Bhatkande et al. (May 2016) applied the Choice Tree Algorithm for trashing the unacceptable data. Generally, maximum heat and minimal heat are primarily responsible for the weather prediction. On the proportion of these variables, believed there is the total cool or whole warm or snowdrop and produced a type using choice pine to believed temperature phenomena like whole cool, whole warm and snowdrop that will be frequently a living keeping information. Neha Khan et al. (Mar 2015) reduced chance in potentially useful information may be extracted from the emotions of the message. Major Information has mainly three traits namely Speed, Quantity, and Range, on the basis of these traits data may be classified in three ways - supervised, unsupervised and partial supervised methods. Numerous formulas and methods are recently planned for Clustering and Classification of the information and E-document. In our record, we will discuss & assess widely used major methods in big data classification. Quande Qin et al. (Mar 2015) ABC algorithm is a simple and easy adjustment to the design of the essential ABC algorithm. Furthermore, the proposed strategy is normal and could be incorporated in different ABC variants. Some 21 standard operates in 30 and 50 measurements are employed in the fresh studies. The fresh benefits show the potency of the proposed time-varying strategy. Hardeep Kaur (Apr 2015) There are numerous programs of data mining in a number of fields like knowledge, scientific and engineering, healthcare, business and other and discussed principles of instructional information mining. As a paper, we shall mostly concentrate on the programs of data mining in the field of education. Sam Fletcher et al. (June 2015) Compared Forest Framework to its precursor, Framework, and another established approach, GDP. Our comparison is completed using our three evaluation standards, besides Forecast Accuracy. Our fresh effects display the achievement your planned extensions to Framework as well as performance your evaluation criteria. Yash Ja-in et al. (Aug 2015) Data mining methods are fast rising and used by a number of analysts in the past few years. The leading advantageous asset utilizing data mining is usually to extracts the results from lots of data which shops in a number of data repository. Knowledge is indicating the significance through primary or indirect. Included the research into numerous data mining techniques. We can even discuss the applying parts and its particular potential scope. A.Yurtkuran et al. (Sept 2015) Applied a fresh option acceptance rule along with a probabilistic multi search method, the intensification, and diversification efficiency within the ABC algorithm is improved. The planned algorithm continues to be tried on well-known benchmark features of countless dimensions by comparing against book ABC variations, as well as several new state-of-the-art algorithms. Computational effects reveal that planned ABC-SA outperforms different ABC variations and is much more advanced than state-of-the-art methods planned through the literature. Ranshul Chaudhary et al. (Jan 2014) Compared with different DM methods, Sensible Programs (ISS) centered strategies, such as Synthetic Neural Communities (ANNs), unclear set theory, approximate reasoning, and derivative-free optimization techniques like Genetic Methods (GAs), are resistant of imprecision, uncertainty, incomplete reality, an approximation. This paper is worried about a few ideas behind design; implementation, screening, and program of a post-ISs centered DM technique. Aarti Sharma et al. (Feb 2014) Data mining is a solid and a fresh field having various techniques. It turns natural data into helpful data in several research fields. It can benefit to discover the patterns to determine potential styles in the medical field. Rohit Pitre et al. (Apr 2014) Huge Results are a new expression accustomed to seeing the datasets that due to their big measurement and complexity. Huge Data are now actually fast growing in all science and engineering domains, including bodily, natural and biomedical sciences. Huge Data mining is the capacity of removing helpful data from all of these big data sets or revenues of data, that simply because of its size, variability, and velocity, it wasn't possible before to perform it. The Huge Data challenge is becoming practically the most interesting possibilities for years. This examine paper is the more knowledge about what is large data, Data mining, Data mining with large data, Demanding problems and its particular connected work.

III. DATA MINING TECHNIQUES

Relevant details should be given including experimental design and the technique (s) used along with appropriate statistical methods used clearly along with the year of experimentation (field and laboratory). Use of records mining for software is primarily based on a number of strategies. A few examples are:

A. Clustering algorithm

Clustering set of rules may be taken into consideration the most essential unsupervised getting to know the problem; A cluster is a group of objects which are "similar" among them and are "varied" to the items belonging to other clusters. We will show this with an easy graphical instance: Use of data

mining for application is based on a number of Techniques. Some examples are:

1. **Medical Field:** In the medical area, the cluster analysis provides a systematic, formalized method for information exploration. It is defining businesses with clinical similarities. Clustering equipment may be used as a proof based totally medicinal drug analysis gadget that could assist in the prevention of health center faults. In addition, they can be green clustering tools reduce call for on pricey healthcare resources. It may assist physician's address the records overload and might assist in destiny making plans for these offerings. Clustering outcomes are was once correlation between diseases and for better insight into scientific survey statistics. These kinds of blessings stimulated the researcher to broaden clustering fashions for scientific statistics.

2. **Records overload:** To grow the number of facts accrued and saved in the healthcare enterprise. Knowledge discovery and retrieval of statistics from huge databases is tough and are prohibitively highly-priced. Too many sickness markers are available for decision making. The excessive awareness for great care among public and accelerated life expectancy is increasing the demand for pleasant fitness services. But with overworked and worn-out physicians, disturbing work conditions, and so on.

B. Time series Algorithm

A Time arrangement is a chain of realities listed in the time arranged. A Time arrangement is a chain taken at progressive also separated focuses in time. Cases of time accumulation are statues of sea tides, tallies of sunspots. Time accumulation is frequently plotted in line outlines. Time arrangement are utilized as a part of insights, sign handling, design ubiquity, econometrics, scientific back, atmosphere estimating, wise transport and way gauging, quake forecast, electroencephalography, control building, space science, correspondences building, and an extensive part in any territory of completed innovative know-how and building which incorporates worldly estimations.

1. **Affected Person Danger Stratification For Health Center:** Time-collection statistics are available in lots of distinct fields, inclusive of medicine, finance, statistics recuperation and climate forecast. Sizable research has been committed to the analysis and class of such alerts. In latest years, researchers have had top-notch achievement with figuring out temporal designs in the time series and with methods that forecast the value of variables. In maximum applications, there is an express time series, e.g., ECG indicators, stock costs, audio recordings, and day by day common temperatures. It considers a singular application of time-collection evaluation, patient hazard. The patient threat has an inherent temporal thing it evolves over the years as it's far encouraged with the aid of intrinsic and extrinsic elements. It far identifies the hospitalized sufferers for excessive risk consequences as a time collection institution task. They endorse and motivate the examine of affected person threat approaches to version the evolution of anger over the direction of a health center admission.

C. Logistic Regression

Logistic regression is the proper regression investigation to lead when the setup factor is paired. All relapse examinations; the strategic relapse is a prescient investigation. Calculated relapse is utilized to disclose information and to clarify the association between one ward twofold factor and one or more noteworthy proportion sans scale factors.

1. **Expect Patron Maintenance:** Logistic regression is a completely effective device in constructing fashions. SAS system software program implements the device in each Proc Logistic inside the SAS/STAT module and the impending enterprise Miner product. While implementing the logistic regression fashions can produce powerful intuitions into why clients are departed and others stay. These insights can then be hired to alter organizational strategies and/or verify the effect of the implementation of those strategies. Any other shape of statistical modeling, to attend to the construct variables signifying the portent of interest and to choose applicant loose variables that are "substantively applicable" to the problems under analysis. Ultimately, the patron choice suggests that periodic re-assessment of the chosen version is exceptionally appropriate. As customer demands and wishes, as well as the surroundings" in which the employer competes, adjustments, so need to the models the corporation employs to predict purchaser retention.

D. Naive Bayes

Naive Bayes is an easy approach. Bayes classifiers anticipate how the fee of a perfect feature depends on the cost of almost every other feature, given the category variable. As an illustration, a fruit may be considered to become a Mango should it be far yellow, spherical, and approximately 10 cm in diameter. A naive Bayes classifier considers every one of these features to contribute independently to the prospect that fruit is usually a Mango, it doesn't matter any possible correlations among large, roundness and diameter functions. In plenty of applications, parameter estimation for naive Bayes fashions uses the technique of maximum likelihood; in other phrases, you will paintings while using the naive Bayes model without accepting Bayesian possibility or the use of any Bayesian techniques.

IV. GAPS IN LITERATURE WORK

The most of the existing techniques have certain constraints because it has neglected many things some of them are:

- The use of an ensemble of data mining techniques can be done to improve the accuracy rate further for brain tumor classification.
- Most of the existing techniques are limited to some significant features of a brain tumor.
- The ensemble of random forest and Artificial Neural Network (ANN) has not been used extensively to improve the accuracy rate further for brain tumor classification.
- The accuracy rate of the existing methods is found to be poor so improvement is required to make them more consistent.

V. CONCLUSION AND FUTURE SCOPE

The Data mining is an important technique. It is used in technology fields such as scientific discovery, banking, insurance, decision support and customer relationship management etc. In the current business environment the salesperson motivation for mass marketing for the promotion of any product. But this way of marketing will not give a successful result to need and preference of the customer. In this, a marketer can perform one to one marketing instead of mass marketing. Since as per our assumption if the effort is applied in a particular direction, the sale will increase and this will result in increased profit which is the ultimate goal of the marketer.

ACKNOWLEDGMENT *(HEADING 5)*

I would like to thank Mrs. Jasleen kaur my faculty guide for the thesis, for her regular guidance without which my thesis would not have been completed. I would also like to thank Head of Department of computer science engineering, Er. Kirandeep Singh for his valuable support and guidance.

REFERENCES

[1] Sam Fletcher et al. "An anonymization technique using intersected decision trees" Journal of King Saud University – Computer and Information Sciences (2015) 27, 297–304.

[2] Ashish Kumar et al. "Implementation and Comparison of Decision Tree Based Algorithms", International Journal of Innovations & Advancement in Computer Science IJIACS ISSN 2347 – 8616 Volume 4, Special Issue May 2015.

[3] Yash Jain et al."A Survey On Data Mining Techniques, Their Application And Future Scope", International Journal Of Engineering Sciences & Research Technology [Jain*, 4.(8): April 2015] ISSN: 2277-9655 (I2OR), Publication Impact Factor: 3.785

[4] Neha Khan et al. "Big Data Classification using Evolutionary Techniques: A Survey",2015 IEEE International Conference on Engineering and Technology (ICETECH), 20th March 2015, Coimbatore, TN, India.

[5] Sonia Singh et al. "COMPARATIVE STUDY ID3, CART AND C4.5 DECISION TREE ALGORITHM: A SURVEY", International Journal of Advanced Information Science and Technology (IJAIST), ISSN: 2319:2682 Vol.27, No.27, July 2014.

[6] Manpreet Singh et al. "Performance Analysis of Decision Trees", International Journal of Computer Applications (0975 – 8887) Volume 71– No.19, June 2013

[7] T.Miranda Lakshmi et al. "An Analysis On Performance Of Decision Tree Algorithms Using Student's Qualitative Data", I.J.Modern Education And Computer Science, 2013, 5, 18-27 Published Online June 2013 In MECS (Http://Www.Mecs-Press.Org/) DOI: 10.5815/Ijmecs.2013.05.03

[8] Nikita Jain, Vishal Srivastava, M. Tech. Scholar, Associate Professor, Arya College Of Engineering And IT, "Data Mining Techniques: A Survey Paper", International Journal Of Research In Engineering And Technology Eisen: 2319-1163 | ISSN: 2321-7308.

[9] Shikha Chourasia, "Survey paper on improved methods of ID3 decision tree classification",

[10] International Journal of Scientific and Research Publications, Volume 3, Issue 12, December 2013 ISSN 2250-3153

[11] Mohd Afizi Mohd Shukran, Faculty of Science & Defense Technology, National Defense University of Malaysia (NDUM) "Artificial Bee Colony based Data Mining Algorithms for Classification Tasks", Canadian Center of Science and Education www.ccsenet.org/mas Modern Applied Science Vol. 5, No. 4; August 2011.

[12] Mrs. Bharati M. Ramageri, Lecturer Of Modern Institute Of Information Technology And Research, Department Of Computer Application, "Data Mining Techniques And Applications", Indian Journal Of Computer Science And Engineering Vol. 1 No. 4 301-305.

[13] Fahad S. Abu-Mouti and Mohamed E. El-Hawary, "Overview of Artificial Bee Colony (ABC) Algorithm and Its Applications" IEEE, 2012.

[14] Er. Ankit Choubey And Dr. G. L. Prajapati, " An Understanding Of Abc Algorithm And Its Applications" International Journal Of Current Engineering And Scientific Research (IJCESR), ISSN (Print): 2393-8374, (Online): 2394-0697, Volume-2, Issue-10, 2015.

[15] Alkin Yurtkuran and Erdal Emel, "An Enhanced Artificial Bee Colony Algorithm with Solution Acceptance Rule and Probabilistic Multisearch" Hindawi Publishing Corporation Computational Intelligence and Neuroscience Volume 2016, Article ID 8085953,2016.

[16] Quande Qin, Shi Cheng, Qingyu Zhang, Li Li and Yuhui Shi, "Artificial Bee Colony Algorithm with Time-Varying Strategy" Hindawi Publishing Corporation Discrete Dynamics in Nature and Society Volume 2015, Article ID 674595, 2015.

[17] Dhanya P Varghese & Tintu P B, "A SURVEY ON HEALTH DATA USING DATA MINING TECHNIQUES", International Research Journal of Engineering and Technology (JET), Volume: 02 Issue: 07, Oct-2015.

[18] Doron Shalvi & Nicholas DeClaris, "AN UNSUPERVISED NEURAL NETWORK APPROACH TO MEDICAL DATA MINING TECHNIQUES", IEEE, 1998.

[19] Gustavo Santos-Garcia & Gonzalo Varela & Nuria Novoa & Marcelo F. Jimenez, "PREDICTION OF POSTOPERATIVE MORBIDITY AFTER LUNG RESECTION USING AN ARTIFICIAL NEURAL NETWORK ENSEMBLE", Artificial Intelligence in Medicine 30:61–69, 2004.

[20] Hojin Moon & Hongshik Ahn & Ralph Kodell & Songjoon Baek & Chien- Ju Lin & James Chen, "ENSEMBLE METHODS FOR CLASSIFICATION OF PATIENTS FOR PERSONALIZED

MEDICINE WITH HIGH-DIMENSIONAL DATA". Artificial Intelligence in Medicine 41:197–207, 2007.

[21] I. Curiac & G. Vasile & O. Banias & C. Volosencu & A. Albu, "BAYESIAN NETWORK MODEL FOR DIAGNOSIS OF PSYCHIATRIC DISEASES", Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat, Croatia, 22-25 June-2009.

[22] Jeong-Yon Shim & Lei Xu, "MEDICAL DATA MINING MODEL FOR ORIENTAL MEDICINE VIA BYY BINARY INDEPENDENT FACTOR ANALYSIS", IEEE.P1-4, 2003.

[23] K.Sharmila & Dr.S.A.Vethamanickam, "SURVEY ON DATA MINING ALGORITHM AND ITS APPLICATION IN HEALTHCARE SECTOR USING HADOOP PLATFORM", International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, Volume: 05, Issue: 01, January-2015.

[24] Andrii Shalaginov, Lars Strande Grini, Katrin Franke (2016) "Understanding Neuro-Fuzzy on a class of multinomial malware detection problems, In Neural Networks (IJCNN)," 2016 International Joint Conference

[25] Huda, Shamsul, et al. "A hybrid feature selection with ensemble classification for imbalanced healthcare data: A case study for brain tumor diagnosis." IEEE Access 4 (2016): 9145-9154.

[26] Azad, S., Fattah, S. A., & Shahnaz, C. (2017, November). "An automatic scheme for brain tumor region detection from 3D MRI data based on enhanced intensity variation." In Region 10 Conference, TENCON 2017-2017 IEEE (pp. 1-6). IEEE.

[27] Ramani RG, Sivaselvi K. Classification of "Pathological Magnetic Resonance Images of Brain Using Data Mining Techniques. In Recent Trends and Challenges in Computational Models (ICRTCCM)," 2017 Second International Conference on 2017 Feb 3 (pp. 77-82). IEEE.

Authors Profile

Jasleen kaur received the B.Tech degree in computer science and engineering from GNDU(Guru Nanak Dev university) in 2009 and M.Tech in computer science and engineering from GNDU(Guru Nanak Dev university) in 2011.She is pursing currently Ph.D* and currently working as Assit Profersor in computer science and five year experience in KCET(Khalsa College Engineering & Technology)

