International Journal of Research (IJR)
p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 5
May 2021

Survey on Data mining approach and Feature Scope

Ayush Soni*
Department of Computer Science, Laxmi Narayan college of technology, Indore
Guided by- Khushboo Sawant*
Department of Computer Science, Laxmi Narayan college of technology, Indore

Abstract— This paper puts forward the 8 most used data mining algorithms used in the research field which are: C4.5, k-Means, SVM, EM, PageRank, Apriori, kNN and CART. With each algorithm, a basic explanation is given with a real time example, and each algorithms pros and cons are weighed individually. These algorithms are seen in some of the most important topics in data mining research and development such as classification, clustering, statistical learning, association analysis, and link mining. To analyze, manage and make a decision of such type of huge amount of data we need techniques called the data mining which will transforming in many fields. This paper imparts more number of applications of the data mining and also o focuses scope of the data mining which will helpful in the further research.

1. INTRODUCTION Data is produced in such large amounts that today the need to analyze and understand this data is of the essence. The grouping of data is achieved by clustering algorithms and can then further be analyzed by mathematicians as well as by big data analysis methods. This clustering of data has seen a wide scale use in social network analysis, market research, medical imaging etc. This grouping of data is seen in many different graphical forms such as the one shown below.

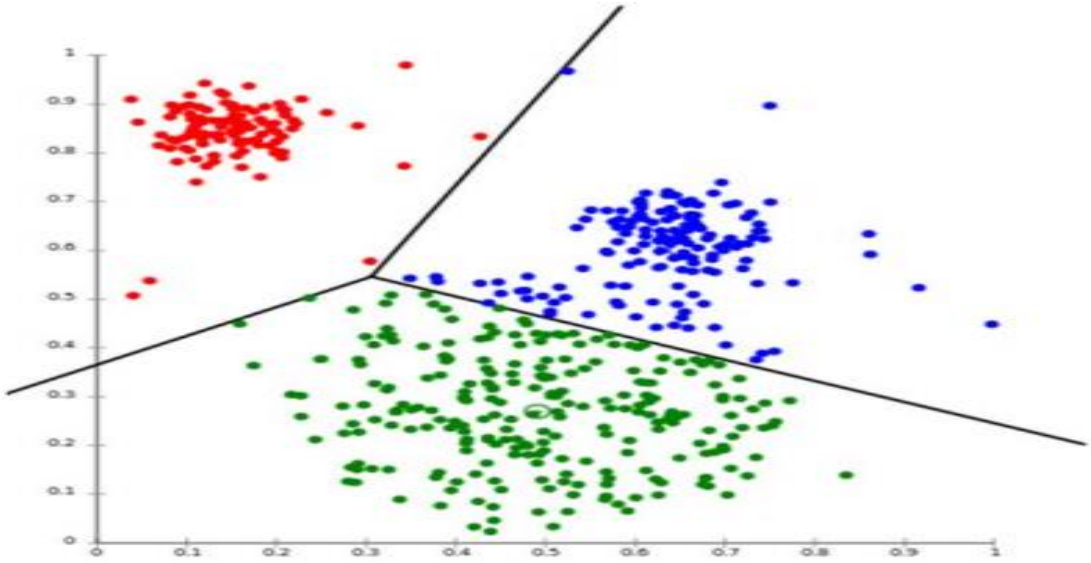

Figure 1 : K-MEANS

Here we have taken an instance to better understand and identify the most influential algorithms that have been widely used in data mining. Most of these were identified during the ICDM '06 in Hong Kong. We deal with a wide variety of algorithms such as clustering, classification, link mining, association analysis and statistical learning. We have analyzed these algorithms in depth and have put forward a simple explanation of these concepts with real world examples to help in the better understanding of the chosen algorithms and have also weighed in each one's pros and cons individually to help with implementation of the algorithms.

2. The Data Mining Task

The data mining tasks are of different types depending on the use of data mining result the data mining tasks are classified as[1,2]: 2.1 Exploratory Data Analysis: In the repositories vast amount of information's are available .This data mining task will serve the two purposes (i).With out the knowledge for what the customer is searching, then (ii) It analyze the data These techniques are interactive and visual to the customer. 2.2 Descriptive Modeling: It describe all the data, it includes models for overall probability distribution of the data, partitioning of the p-dimensional space into groups and models describing the relationships between the variables.

2.3 Predictive Modeling: This model permits the value of one variable to be predicted from the known values of other variables.

2.4. Discovering Patterns and Rules: This task is primarily used to find the hidden pattern as well as to discover the pattern in the cluster. In a cluster a number of patterns of different size and clusters are available .The aim of this task is "how best we will detect the patterns" .This can be accomplished by using rule induction and many more techniques in the data mining algorithm like(K-Means /K-Medoids) .These are called the clustering algorithm.

2.5 Retrieval by Content: The primary objective of this task is to find the data sets of frequently used in the for audio/video as well as images It is finding pattern similar to the pattern of interest in the data set

3. Types of Data Mining System: Data mining systems can be categorized according to various criteria the classification is as follows[3]:

3.1 Classification of data mining systems according to the type of data source mined: In an organization a huge amount of data's are available where we need to classify these data but these are available most of times in a similar fashion. we need to classify these data according to its type(maybe audio/video ,text format etc)

3.2 Classification of data mining systems according to the data model: There are so many number of data mining models (Relational data model, Object Model, Object Oriented data Model,

Hierarchical data Model/W data model)are available and each and every model we are using the different data .According to these data model the data mining system classify the data in the model.

3.3 Classification of data mining systems according to the kind of knowledge discovered: This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

3.4 Classification of data mining systems according to mining techniques used: This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems. A comprehensive system would provide a wide variety of data mining techniques to fit different situations and options, and offer different degrees of user interaction.

4. Data Mining Life Cycle: The life cycle of a data mining project consists of six phases[2,4]. The sequence of the phases is not rigid. Moving back and forth between different phases is always required. It depends on the outcome of each phase The main phases are:

4.1. Business Understanding: This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this knowledge into a data mining problem definition and a preliminary plan designed to achieve the objectives.

4.2 Data Understanding: It starts with an initial data collection, to get familiar with the data, to identify data quality problems, to discover first insights into the data or to detect interesting subsets to form hypotheses for hidden information.

4.3 Data Preparation: In this stage , it collects all the different data sets and construct the varieties of the activities basing on the initial raw data

4.4 Modeling: In this phase, various modeling techniques are selected and applied and their parameters are calibrated to optimal values.

4.5 Evaluation: In this stage the model is thoroughly evaluated and reviewed. The steps executed to construct the model to be certain it properly achieves the business objectives. At the end of this phase, a decision on the use of the data mining results should be reached.

4.6 Deployment: The purpose of the model is to increase knowledge of the data, the knowledge gained will need to be organized and presented in a way that the customer can use it. The deployment phase can be as simple as generating a report or as complex as implementing a repeatable data mining process across the enterprise.

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 5
May 2021

5. Apriori : It is applied to a dataset containing a large number of transactions. The algorithm learns associate rules. In data mining, associate rules are techniques for learning relations and correlations among variables in database. Example for Apriori algorithm: Consider the dataset of a supermarket transaction to be a giant spreadsheet. Each row of the spreadsheet is a customer transaction and every column represents a different grocery item. By using Apriori algorithm, we can analyze the items that are purchased together. We can also find the items that are frequently purchased than the other items. Together, these items are called item sets.

Table 1: Apriori Example:

Transaction ID	Chips	Dip	Soda	Apples	Milk
1	X	X	X		
2	X	X			X
3	X		X		

The main aim of this is to make the shoppers buy more. For example: You can see that chips & dip and chips & soda seem to frequently occur together. These are called 2- itemsets. With a large enough dataset, it is much harder to "see" the relationships especially when you're dealing with 3-itemsets or more. Working of Apriori: three things need to be defined before starting with the algorithm. They are: The size of the item set (If you want to see the patterns for 2-itemset, 3-itemset, etc.?), the number of transactions containing the item set divided by the total number of transactions. A frequent item set is one which meets the support and Confidence or conditional probability. Apriori algorithm has 3 steps of approach: Join, Prune and Repeat. Apriori is an unsupervised learning approach. It discovers or mines for interesting patters and relationships. Apriori can also be supervised to do classification on labeled data. Apriori can be used for ARtool, Weka, and Orange. The advantages are: it uses large item set property, easily Parallelized, easy to implement. The limitations are: Assumes Transaction database is memory resident and requires many database scans.

6. Page rank : It is a link algorithm used to determine the relative significance of certain object linked within a network of objects. Link analysis is similar to network analysis which is looking to search the association among links. The most common example for page rank is Google's search engine. Though, the search engine doesn't solely rely on page rank. It's one of the measures goggle uses to determine a web page importance. Advantages are: Robust against spam, global measure and query independent. The limitations are: Favors older pages.

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 5
May 2021

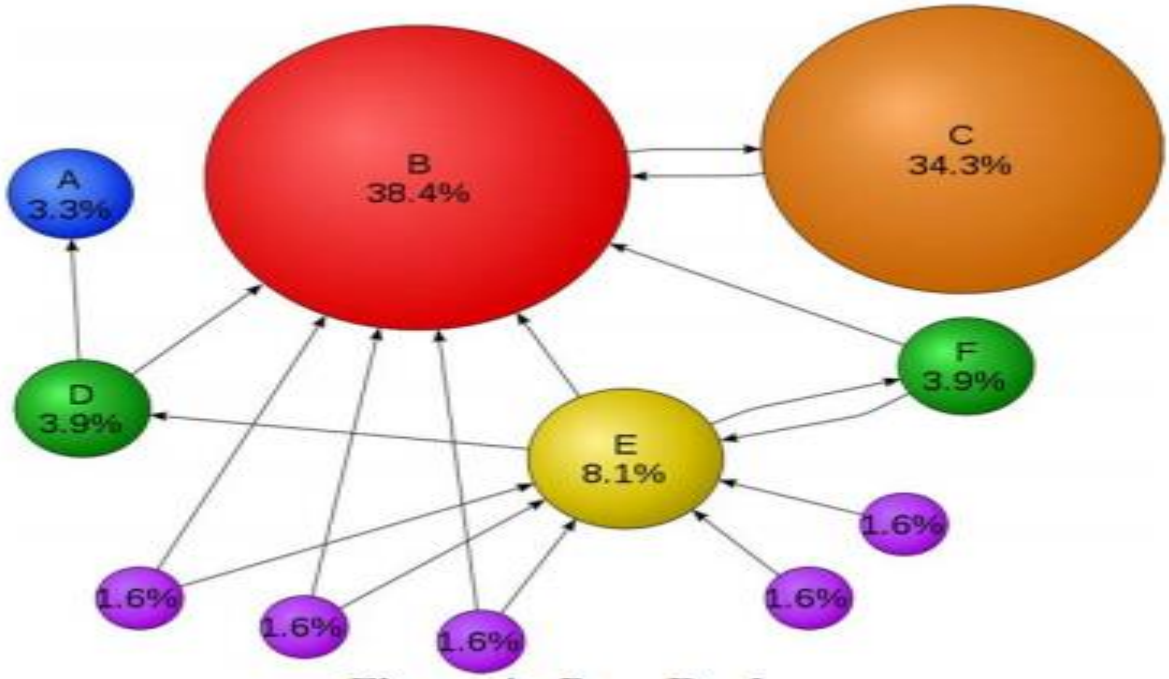

Figure 4 : Page Rank

7. Data Mining Methods:

Some of the popular data mining methods are as follows:

1 Decision Trees and Rules

2 Nonlinear Regression and Classification Methods

3 Example-based Methods

4 Probabilistic Graphical Dependency Models

5 Relational Learning Models

We found these are some famous data mining methods are broadly classified as: On-Line Analytical Processing ,(OLAP), Classification, Clustering, Association Rule Mining, Temporal Data Mining, Time Series Analysis, Spatial Mining, Web Mining etc. These methods use different types of algorithms and data. The data source can be data warehouse, database, flat file or text file. The algorithms may be Statistical Algorithms, Decision Tree based, Nearest Neighbor, Neural Network based, Genetic Algorithms based, Ruled based, Support Vector Machine etc. Generally the data mining algorithms are fully dependent of the two factors these are

(i) which type of data sets are using

(ii) what type of requirements of the user

Basing upon the above two factors the data mining algorithms are used.A knowledge discovery (KD) process involves preprocessing data, choosing a data-mining algorithm, and post processing the mining results. The Intelligent Discovery Assistants [7] (IDA), helps users in applying valid knowledge discovery processes. The IDA can provide users with three benefits:

(i) A systematic enumeration of valid knowledge discovery processes;
(ii) Effective rankings of valid processes by different criteria, which help to choose between the options;
(iii) An infrastructure for sharing knowledge, which leads to network externalities.

Several other attempts have been made to automate this process and design of a generalized data mining tool that posse's intelligence to select the data and data mining algorithms and up to some extent the knowledge discovery.

8. Data Mining Applications: In this section, we have focused some of the applications of data mining and its techniques are analyzed respectively Order.

1 Data Mining Applications in Healthcare Data mining applications in health can have tremendous potential and usefulness [60]. However, the success of healthcare data mining hinges on the availability of clean healthcare data. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications

Future Directions of Health care system through Data Mining Tools As healthcare data are not limited to just quantitative data (e.g., doctor's notes or clinical records), it is necessary to also explore the use of text mining to expand the scope and nature of what healthcare data mining can currently do. This is specially used to mixed all the data and then mining the text. It is also useful to look into how images (e.g., MRI scans) can be brought into healthcare data mining applications. It is noted that progress has been made in these areas.

The Scope of Data Mining Data mining derives its name from the similarities between searching for valuable business information in a large database for example, finding linked products in gigabytes of store scanner data and mining a mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides. Given databases of sufficient size and quality, data mining technology can generate new business opportunities by providing these capabilities:

p-ISSN: 2348-6848
e-ISSN: 2348-795X
Vol. 8 Issue 5
May 2021

Conclusion: In this paper we briefly reviewed the various data mining applications. This review would be helpful to researchers to focus on the various issues of data mining. In future course, we will review the various classification algorithms and significance of evolutionary computing (genetic programming) approach in designing of efficient classification algorithms for data mining. Most of the previous studies on data mining applications in various fields use the variety of data types range from text to images and stores in variety of databases and data structures. The different methods of data mining are used to extract the patterns and thus the knowledge from this variety database. Selection of data and methods for data mining is an important task in this process and needs the knowledge of the domain. Several attempts have been made to design and develop the generic data mining system but no system found completely generic. Thus, for every domain the domain expert's assistant is mandatory. The domain experts shall be guided by the system to effectively apply their knowledge for the use of data mining systems to generate required knowledge. The domain experts are required to determine the variety of data that should be collected in the specific problem domain, selection of specific data for data mining, cleaning and transformation of data, extracting patterns for knowledge generation and finally interpretation of the patterns and knowledge generation. Most of the domain specific data mining applications show accuracy above 90%. The generic data mining applications are having the limitations. From the study of various data mining applications it is observed that, no application called generic application is 100 % generic. The intelligent interfaces and intelligent agents up to some extent make the application generic but have limitations. The domain experts play important role in the different stages of data mining. The decisions at different stages are influenced by the factors like domain and data details, aim of the data mining, and the context parameters. The domain specific applications are aimed to extract specific knowledge. The domain experts by considering the user's requirements and other context parameters guide the system. The results yield from the domain specific applications are more accurate and useful. Therefore it is conclude that the domain specific applications are more specific for data mining. From above study it seems very difficult to design and develop a data mining system, which can work dynamically for any domain.

REFERENCES

[1] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Two Crows Corporation, 10500 Falls Road, Potomac, MD 20854 (U.S.A.), 1999.

[2] Larose, D. T., "Discovering Knowledge in Data: An Introduction to Data Mining", ISBN 0-471-66657-2, ohn Wiley & Sons, Inc, 2005.

[3] Dunham, M. H., Sridhar S., "Data Mining: Introductory and Advanced Topics", Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition, 2006

[4] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. and Wirth, R... "CRISP-DM 1.0 : Step-by-step data mining guide, NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringenen Bank Group B.V (The Netherlands), 2000".

[5] Fayyad, U., Piatetsky-Shapiro, G., and Smyth P., "From Data Mining to Knowledge Discovery in Databases," AI Magazine, American Association for Artificial Intelligence, 1996.

[6] Tan Pang-Ning, Steinbach, M., Vipin Kumar. "Introduction to Data Mining", Pearson Education, New Delhi, ISBN: 978-81-317-1472-0, 3rd Edition, 2009. Bernstein, A. and Provost, F., "An Intelligent Assistant for the Knowledge Discovery Process", Working Paper of the Center for Digital Economy Research, New York University and also presented at the IJCAI 2001 Workshop on Wrappers for Performance Enhancement in Knowledge Discovery in Databases.

[7] Baazaoui, Z., H., Faiz, S., and Ben Ghezala, H., "A Framework for Data Mining Based Multi-Agent: An Application to Spatial Data, volume 5, ISSN 1307-6884," Proceedings of World Academy of Science, Engineering and Technology, April 2005.

[8] Rantzau, R. and Schwarz, H., "A Multi-Tier Architecture for High-Performance Data Mining,A Technical Project Report of ESPRIT project, The consortium of CRITIKAL project, Attar Software Ltd. (UK), Gehe AG (Denmark); Lloyds TSB Group (UK), Parallel Applications Centre, University of Southampton (UK), BWI, University of Stuttgart (Denmark), IPVR, University of Stuttgart (Denmark)".

[9] Botia, J. A., Garijo, M. y Velasco, J. R., Skarmeta, A. F., "A Generic Data mining System basic design and implementation guidelines", A Technical Project Report of

[10] Campos, M. M., Stengard, P. J., Boriana, L. M., "Data-Centric Automated Data Mining" WebSite.:
www.oracle.com/technology/products/bi/odm/pdf/automated_data_mining_paper_1205.pdf

[11] Sirgo, J., Lopez, A., Janez, R., Blanco, R., Abajo, N., Tarrio, M., Perez, R., "A Data Mining Engine based on Internet, Emerging Technologies and Factory Automation," Proceedings ETFA '03, IEEEz Conference,16-19Sept.2003.WebSite:www.citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.11.8955

[12] Bianca V. D.,Philippe Boula de Mareüil and Martine Adda-Decker, "Identification of foreign-accented French using data mining techniques, Computer Sciences Laboratory for

Mechanics and Engineering Sciences (LIMSI)".Website www.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf

[13] Bianca V. D.,Philippe Boula de Mareüil and Martine Adda-Decker, "Identification of foreign-accented French using data mining techniques, Computer Sciences Laboratory for Mechanics and Engineering Sciences (LIMSI)".Website www.limsi.fr/Individu/bianca/article/Vieru&Boula&Madda_ParaLing07.pdf

[14] Halteren, H. V., Oostdijk N., "Linguistic profiling of texts for the purpose of language verification, The ILK research group, Tilburg centre for Creative Computing and the Department of Communication and Information Sciences of the Faculty of Humanities, TilburgUniversity,TheNetherlands."Website: www.ilk.uvt.nl/~antalb/textmining/LingProfColingDef.pdf

[15] Antonie, M. L., Zaiane, O. R.,Coman, A., "Application of Data Mining Techniques for Medical Image Classification", Proceedings of the Second International Workshop on Multimedia Data Mining MDM/KDD 2001) in conjunction with ACM SIGKDD conference, San Francisco, August 26, 2001.