International Journal of Advances in Engineering and Management (IJAEM)
Volume 4, Issue 5 May 2022, pp: 1118-1123 www.ijaem.net ISSN: 2395-5252

An In-Depth Study of the Different Data Mining and Machine Learning Techniques

Satish Sahani[1], Neha Gond[2], Dr. Harvendra Kumar[3]
[1]*B.Tech 4th Year, Dept. of Information Technology and Engineering, ITM Gorakhpur, U.P, India*
[2]*B.Tech 4th Year, Dept. of Information Technology and Engineering, ITM Gorakhpur, U.P, India*
[3]*Associate Professor, Dept. of Computer Science and Engineering, ITM Gorakhpur, U.P, India*

Submitted: 10-05-2022 Revised: 17-05-2022 Accepted: 20-05-2022

ABSTRACT: Data mining is the process of extracting meaningful patterns and information from large amounts of data. It is a fast expanding research topic, whose appeal and demand have risen in tandem with the advent of tools that may assist in revealing and comprehending hidden information in massive amounts of data. Several entities, such as banks, insurance firms, and grocery stores, produce such data on a daily basis. This paper describes the various DM approaches in depth, with illustrations, and compares them in terms of performance parameters such as execution time and efficiency. This paper covers the following sections: Introduction, Data Mining, Data Mining Model, Performance Measure and Conclusion.

Keyword - Data Mining, Data mining Model, Performance Measure, Association Rule Mining.

I. INTRODUCTION

The growing interest in DM and the use of historical data to find irregularities that refine future decisions follows the convergence of some recent trends, like the falling amount of large facts storage gadgets and the growing ease of assembling data over networks. The enlargement of robust and well-organised (ML) algorithms to operate this data and the falling cost of calculation power are enabling the utilisation of computationally exhaustive and intensive data analysis techniques. Data mining is the practise of identifying well-organized, new, potentially helpful, and intelligible patterns in data. With the worldwide utilisation of databases and the explosive widening of their dimensions, organisations are looking towards the difficulty of facts overload. The issue of effectively using these huge volumes of data is becoming a major problem for all enterprises. As a result, while some techniques such as clustering, decision trees, and so on exist in many fields, they are not well suited to DM for association rule mining. As a result, in order to use statistical and mathematical tools, we must adapt these techniques so that they can efficiently sift through enormous amounts of data stored in secondary memory.

II. MACHINE LEARNING

As stated by Arthur Samuel, machine learning is defined as the research-based study of algorithms and analytical models that provide computers with the capability to learn in the absence of explicitly programmed instructions. It's used to trained machine how to manage the data even more efficiently, e.g., a concept representing the outcome of learning as output. Inductive learning, where the arrangement infers grasp itself by observing its environment, has three leading strategies: supervised learning, unsupervised learning, and reinforcement learning. Inductive learning methods can be used to forecast the results of future situations; in other words, not only for states encountered but rather for unseen states that could occur. There can be many possible models based on having handsets as examples. In such situations, the principle of Occam's razor is very appropriate. It states that if there are many explanations for a particular phenomenon, it makes sense to choose the simplest one because it is more likely to capture the nature of the phenomenon.

1.1. Supervised Learning

Supervised learning means learning from examples, where an instruction set is given that acts as an example for the classes. The system then finds a description of each category. Once the description (and hence the classification rule) has been formulated, it's used to forecast the class of previously unseen objects. This is analogous to the distinction between analysis and interpretation in statistics. It uses levelled data. Classification is a powerful learning technique. It is a way of categorising a piece of information for already defined groups. The classification method provides mathematical algorithms such as linear programming, decision trees, ANNs, and statistics. For example, classification is applied in an application that makes up for all records of company members who have left or who will probably plan to leave the company in the future. Here, employees may be classified as on leave or staying. Or a bank loan manager analyses his/her customer's data to determine which loan applicants are either safe/risky.

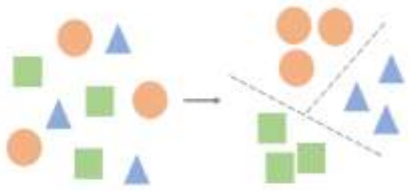


Fig -1: Overview of Supervised Learning (Input examples are categorized into a known set of classes)

1.2. Unsupervised Learning

Unsupervised learning is study from observation and discovery. In this mode of learning, there is no training set or prior knowledge of the classes. The system examines the given set of facts to observe similarities emerging out of the subsets of the facts. The outcome is the acquisition of class descriptions, one for each class, discovered in the environment. Clustering is an example of this technique. Clustering is a technique for grouping similar types of data. Cluster analysis refers to clusters that are similar, but other clusters have different objects. The cluster technique and the classification technique are almost the same kinds of techniques. The classification technique uses predefined classes, while clustering does not use predefined classes. Groups or classes are obtained in clustering by using similarity between records according to the characteristics of real data. The clustering method describes groups and objects in a dataset, while the classification technique is used in predefined classes to assign objects. A book in a library is an example of clustering techniques. The algorithms used in DM and the segment of ML dealing with learning from examples overlap, as does the problem addressed. DM is concerned with discovering understandable knowledge, while ML is concerned with improving the performance of an intelligent system or agent for problem-solving tasks.

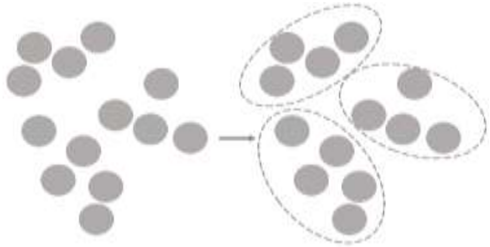


Fig -2: Overview of Unsupervised Learning (Input samples are grouped into clusters based on the underlying patterns)

1.3. Reinforcement Learning

Reinforcement learning is applied when the task at hand is to make a sequence of decisions towards a final reward. During the learning process, an artificial agent gets either rewards or penalties for the actions it performs. Its target is to enhance the total reward. Examples include learning a negotiator to play computer games or perform robotics tasks with the end goal in mind.

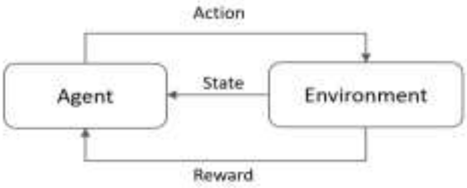


Fig -3: Overview of reinforcement learning (An agent observes the environment state and performs actions to maximize an overall reward)

III. DATA MINING

DM is the procedure of extracting useful information and patterns from vast data. Research in databases and information technology has given rise to an approach to storing and manipulating this precious data for further decision-making. This is an important step in obtaining information from a website that produces useful patterns or models. The goal of this process is to find previously unknown patterns. DM uses statistical algorithms to separate data and evaluate the likelihood of upcoming events. DM is also known as KDD. Example: Mobile service suppliers use DM to outline their marketing campaigns and keep customers from migrating to different vendors. From huge amounts of facts such as payment information, email, text messages, web data shipping, and customer service, DM tools can predict 'churn' by telling customers they want to switch vendors. An opportunity result is provided by these outcomes. Mobile service providers are now free to dispense compensation, and specials to customers at high risk of withdrawal. This type of mining is frequently used by major service suppliers such as broadband, telephone, gas suppliers, etc. DM can be categorised into two dimensions: predictive DM model and descriptive DM model.

3.1 Predictive Model

Predictive-modeling is a mathematical method which aims to forecast future incidents based on past behavior. There are a number of DM functions, such as ARM, Time Series Analysis, Prediction, etc. It works by examining facts to recognise patterns and then using those patterns to create algorithms that data scientists train. It is the fundamental function of predictive analytics applications and is broadly used in a variety of applications, from meteorology to fraud detection, to security management, etc. In short, predictive fashion is a statistical technique using ML and DM to predict and foresee likely future outcomes with the aid of factual and existing data. It acts by analysing current and factual data and projecting what it learns onto a model generated to forecast likely outcomes. This data mining model is broken into four subcategories, namely

Classification:- Classification is a kind of SL that aims to accurately forecast the target category in each case of data. It is the way of DM that assigns objects to a group into specific groups or classes. The principle of separation is to exactly predict the target class in each case of data. It's used to separate each item by a set of data into a predefined set of classes or groups. The division of a data analysis function is where the model or separator is designed for predicting category labels. It allows you to arrange data sets of all sorts, including rigid and huge datasets as well as small and easy ones. The primary aim of classification is to connect a variable of interest with the required variables. For example, a bank loan manager identifies whether the loan applicant is risky or safe.

Regression:- Regression is an approach that is used to predict a range of numeric values (also called continuous values) given in a specific dataset. It refers to a kind of SL that is used to predict some continuous-valued attribute. Regression helps some business institutions analyse the desired variable and predictor variable relations. It is a significant device to analyse the facts that can be used for commercial forecasting and time series modeling. Regression involves the method of fitting a straight line or a curve to numerous data points. It happens in such a way that the distance between the data points and cures comes out to be small. The most popular kinds of regression are linear and logistic regressions. Other than that, many other kinds of regression can be performed depending on their performance on a separate data set.

Time Series Analysis:- In simple words, "a time series is nothing but a sequence of many data points that occur in a successive order for a given period." It represents a series of time-based orders that would be in years, months, weeks, days, hours, minutes, and seconds. It is an observation from the sequence of discrete-time successive intervals. It typically requires a huge number of data points to ensure consistency and reliability. It is a particular way of analysing a series of data points collected over an interval of time. In this analysis, analysts record data points at compatible intervals over a set period rather than recording the data points periodically or randomly.

Prediction:- Prediction is used to observe a numerical result same as in classification, the training dataset contains the inputs and corresponding numerical output values. The algorithm derives the model or a predictor according to the training dataset. The model should observe a numerical output when the new data is given. Unlike in classification, this method does not have a class label. The model predicts a continuous-valued function or sequence value. Predicting the value of a house depending on facts such as the number of rooms, the total area, etc., is an example of prediction.

3.2 Descriptive Model

Non-trivial datasets are studied in the descriptive DM model. It provides information to understand what is going on inside the data. In general, this is the DM model that identifies the patterns or/and relationships in the data. This model uses UL techniques. An example of a descriptive DM model is the retailer. They try to recognise the relationship between the purchased items. On further classification, the predictive model breaks into four subcategories, namely:

Clustering:- Clustering is the way of converting a group of abstract objects into classes of identical objects. It is the procedure of splitting a set of data or objects into a set of significant subclasses called clusters. A subset of objects such that the gap between any of the two objects in the cluster is less than the gap between any object in the cluster and any object that is not located inside it. It helps users to understand the systematic or natural grouping in a data set's aims and to add it either as a stand-alone instrument to get a better perception of data distribution or as a pre-processing step for further algorithms.

Association Rule:- The Association Rule is used to identify frequently used items in a vast data set. It associates the existence of a set of items with another price range in another set of variables. This rule strives to obtain patterns in data based on interpersonal relationships in the same practice. This kind of strategy helps businesses make certain decisions, such as catalogue design, advertising, and customer purchasing behaviour analysis. These are used to discover correlations and co-occurrences between data sets. They are ideally used to describe patterns in data from seemingly unconventional information repositories, such as relational databases and transactional databases. The act of using association rules is sometimes referred to as "association rule mining".

Sequence Discovery:- Sequence discovery, or successive pattern mining, is a DM method that discovers statistically applicable patterns in sequential data. This mining programme assesses definite criteria, such as happening frequency, duration, or values in a set of sequences to find interesting hidden patterns and uncover relationships among data. This technique defines a sequential pattern of events and actions. For example, a customer who buys more than twice in the first phase of the year may be likely to buy at least once during the second quarter.

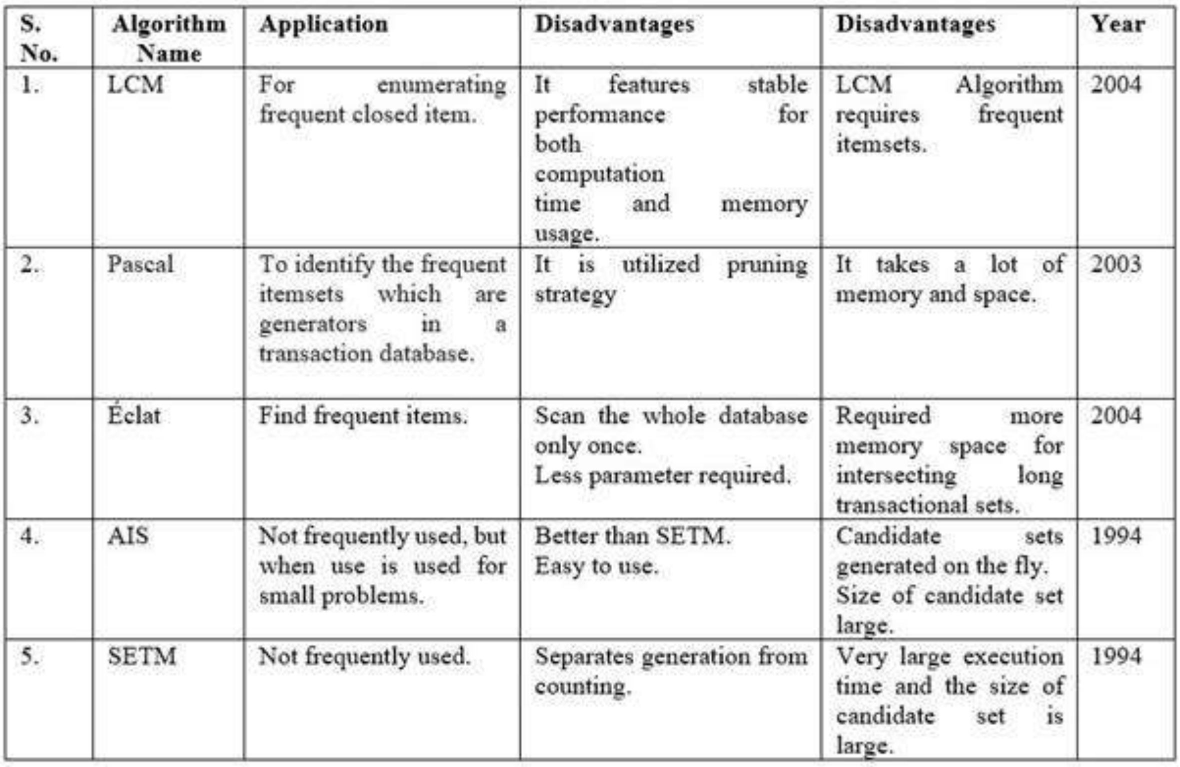
Summarization:- Summarization is a kinds of UL technique that focuses on the compression of data. It is obtained by identifying attributes such as customer name, contact, etc. with many different values. and by removing them or performing a threatening function. Also, we can use standard statistics on the data to represent its summary. This is the key DM concept that involves techniques for finding a compact description of a dataset. Straightforward summarization techniques such as tabulating the mean and standard deviations are often applied for data analysis, data visualization, and automated report generation.

IV. LITERATURE SURVEY

Ting Witten and Yosef Hasan Jbara [1] write about ways to improve efficiency and scalability by executing several learning processes and combining the collective results. The most well-known learning algorithm used in estimating the values of the weights of a neural network —is back propagation, which stores more information during training and then uses this elaborate information to generate better predictions about the test data. Freund and Schapire and Yu and Liu [2], focused on refining and overcoming the classification correctness by scaling down the required classification time. Furthermore, choosing a more suitable distance metric for the specific dataset can improve the correctness of instance-based classifiers. One other choice in designing a training set reduction algorithm is to change the instances using a new representation such as prototypes. They do this independently of all the other classes in the training set. For this reason, for small datasets, it may be better to use a divide-and-conquer algorithm that considers the entire set at once. Villada Drissi and Breiman [3], give the important observation about the issue faced by users while using the frequent patterns of DM. The combining of clustering and classification is known as the prediction analysis clustering algorithm groups the data according to their similarity and the classification algorithm assigns the class to the data. Therefore, the observed problem of data security in the OS will be evaluated and verified against the many parameters set out in the conditions, and it is necessary to access the satisfaction of the above-mentioned conditions. Hassan M. Najadatet al. [4], focused on developing the latest apriori-based algorithm, which satisfies positive aspects of constraints itemsets based mining as anti-monotonicity. The trouble with ARM is that it tries to retrieve matching itemsets for which they are present and for which they represent new constraints, known as relation-based constraints, applicable to the relevant data. In the CIM algorithm, this helps us to recognise the main components of a candidate’s key itemsets and generate frequent itemsets, which satisfies anti-monotonicity properties, which means a little coverage cardinal size restricted to a specific dataset.

R. Srikant et al. [5] proposed the reason for selecting FP-Growth: FP-Growth is a popular, convenient, and efficient enough algorithm for many applications and has replaced LCM, which is faster and uses more memory. It is important to promote SSD quality with an advanced mining algorithm. Significant contributions result in novel results. The support methodology is an efficient and convenient procedure to promote SSD quality. This represents an important contribution for many researchers and companies in vast data analysis. This study is also the first trial with a novel application in terms of using the association rule for SSD quality from an industrial perspective, thereby providing the rationale for its novelty. The study output justifies the capability of the association rule for DM implicational areas. G Nandi and A Das [14] introduced the reversed ARM as one of the most dominant ways to extract confidential information. Aprion was considered the first algorithm for mining organizations. This algorithm is based on complete search and dividing the ARM problem into two sub-problems. The first one is to find all available common sets of data, and the second one is outputs organizational rules from the step of pre-acquired item sets. Then, with the growth of FP, another mining algorithm often overcomes the obstacles of Apriori. Rupam Some [15] propose a variety of ways to hide association rules on the database and to develop support, confidence, and measures. Every method has its own advantages and disadvantages. Its advantage is that it does not produce false rules. The proposed techniques will be compatible with a measurable database. Therefore, the proposed methods of data encryption for the operating system will be evaluated and verified with respect to the many parameters as set out in the conditions.

Table -1 Total Study of Algorithm: [8]



S. No.	Algorithm Name	Application	Disadvantages	Disadvantages	Year
1.	LCM	For enumerating frequent closed item.	It features stable performance for both computation time and memory usage.	LCM Algorithm requires frequent itemsets.	2004
2.	Pascal	To identify the frequent itemsets which are generators in a transaction database.	It is utilized pruning strategy	It takes a lot of memory and space.	2003
3.	Éclat	Find frequent items.	Scan the whole database only once. Less parameter required.	Required more memory space for intersecting long transactional sets.	2004
4.	AIS	Not frequently used, but when use is used for small problems.	Better than SETM. Easy to use.	Candidate sets generated on the fly. Size of candidate set large.	1994
5.	SETM	Not frequently used.	Separates generation from counting.	Very large execution time and the size of candidate set is large.	1994

V. CONCLUSIONS

This paper discussed the various learning techniques and gave a brief overview of various kinds of data mining, and their models. DM is a concept of drawing knowledge or patterns from large amounts of data, and DM can be categorised into two categories based on what a particular project is trying to accomplish. Those two methods are called "descriptive models" and "predictive models." Depending on the type of application, different models could be used to calculate the interesting rules. The study shows that in sequence to achieve the efficiency in DM. It is important to choose an almost acceptable algorithm for the application. The study also showed that more efficient and refined mining methods needed to be developed.

REFERENCES

[1]. Ting Witten and Yosef Hasan Jbara, “Predictingpatterns Using Data Mining Techniques”, Washington DC 20052, vol. 6, 2016.

[2]. Freund et al., “Learning technique: association rule,” Middle-East Journal of scientific research, vol. 5, 2019.

[3]. VilladaDrissi andBreiman, “Learning technique:Mining Association Rules in Student’s Assessment Data”, IJCSI International Journal of Computer Science Issues, vol. 9, Issue 5, 2012.

[4]. Hassan M. Najadat et al., “An Improved Apriori Algorithm for Association Rules”, International Research Journal of Computer Science and Application, vol. 1, 2019.

[5]. R. Srikant et al.,“Algorithms for Mining Association Rules”, 20th International Conference on Very Large Data Bases (VLDB), pp. 487–499, 2018.

[6]. Mrs. Bharati and M. Ramageri, “Data Mining Techniques and Applications”, Indian Journal of Computer Science and Engineering, vol.1, pp. 301-305, 2010.

[7]. ShitalBhojani and Nirav Bhatt, “Data Mining Techniques and Trends”, Indian Journal of Computer Science and Engineering, vol.5, 2016.

[8]. Anshu at al., “Review Paper on Data Mining Techniques and Applications”, International Journal of Innovative Research in Computer Science & Technology (IJIRCST), vo.7, 2019.

[9]. G.Kesavaraj and Dr.S.Sukumaran, “A Study on Classification Techniques in Data Mining”, Fourth International Conference on Computer Communication and Networking Technologies, 2013.

[10]. Kabra. R, Bichkar. R, “Performance Prediction of Engineering Student using Decision Tree”, International Journal of Computer Application December 2011.

[11]. Han. J et al., “Data Mining Concepts and Techniques”, Third Edition the Morgan Kaufmann Series in Data Management Systems Morgan Kaufmann Publishers, 2011.

[12]. Kwame Boakye Agyapong et al., “An Overview of Data Mining Models (Descriptive and Predictive)”, International Journal of Software & Hardware Research in Engineering, vol. 4, 2016.

[13]. B. Nigam et al., “The Comparative Study of Top 10 Algorithm for Association Rule Mining”, International Journal of Computer Sciences and Engineering, vol. 5, 2017.

[14]. G Nandi, A Das, “A Survey of Data Mining Techniques ”, IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 6, No 2, November 2013.

[15]. Rupam Some, “Mining Technique and its Future Trends”, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 6, June 2013.