



## THEORITICAL FRAMEWORK OF DATA MINING, MACHINE LEARNING, ARTIFICIAL INTELLIGENCE AND DATA SCIENCE

**Kalpana Pali**, Scholar, Dr. C. V. Raman University Bilaspur

**Dr. Laxmikant Tiwari**, Assistant Professor, Deptt of Computer Science, Dr. C. V. Raman University Bilaspur

### ABSTRACT

In this paper Data Mining, Machine learning, Artificial Intelligence and Data Science methodologies are explained with all their respective techniques. These techniques systematically segregates large datasets, categorized them based on common traits, predict future events with relevant action and also triggers the actions based on actual events occurred. These four methodologies work independently on defined data sets and provide required inputs to next process. This paper also discusses linkage between these four processes for complete implementation of an Artificial Intelligence based application.

**Keyword:** Data Mining, Machine learning, Artificial Intelligence, Data Science

### INTRODUCTION

In this paper Data Mining, Machine learning, Artificial Intelligence and Data Science techniques are explained. Data mining handles huge and complex database to classify required information to arrange data sets into logical groups. In Machine learning the machine automatically study the factors of models from the given data. It uses self-learning algorithms to improve its outcome with experience over the period. Artificial Intelligence is used to provide answers to the existing problem with the help of data generated by the mining algorithms. Data Science is the most effective technique to identify outliers.

### DATA MINING

Data Mining is used to extract vital information from the dataset. Different data mining algorithms are used to highlight patterns from existing information. This vital information retrieved from data mining algorithms serves as foundation for AI and machine learning. Data mining is more than gathering supersets of facts from different sources. Data mining first retrieve the important information and then aggregate the retrieved information to form a patterns and correlations between data. The designed patterns and correlated data mined data can be used by artificial intelligence and machine learning as the basis. The objective of data mining is to search hidden patterns and relations between two or more data variables from the dataset to generate a commercial value. This data information can be used to predict the results and to summarise the data. There are various data sorting methods which can be used to complete this mining objective such as clustering, classification and sequence analysis. Batched information is used by mining methods to retrieve a new perception for a specific time rather than continuing basis.

#### Data Mining Techniques

Following are the two types of Data Mining Techniques:

- a. **PREDICTIVE MODEL** Predictive model is based on historic data. This model analyses past data and accordingly predicts future events.
  - Classification: Classification is a classic data mining technique based on machine learning. Basically, classification is used to classify each item in a set of data into one of a predefined set of classes or groups. It uses supervised learning.
  - Regression: This process is used to map a data item to a real valued prediction variable.
  - Time Series Analysis: Time Series Analysis examines value of an attributes which varies over a period of time. This analysis inspects the changes in an attribute with periodic intervals.
  - Prediction: This process predicts future state of a variable based on systematic analysis of series of events.



b. **DESCRIPTIVE MODEL:** Descriptive Model identifies patterns or relationships in data. This model is very useful to explore the properties of data.

- **CLUSTERING:** In this process data with similar or matching characteristics are grouped together to form clusters. This is a useful classification technique using large data. K-means is signal processing originated method of vector quantization, which is a famous method for cluster analysis in data mining.
- **SUMMARIZATION:** This process creates subsets of data with associated simple descriptions. This process uses generalization.
- **Association Rules or Link Analysis:** This process examines data to establish relationships among data.
- **Sequence Discovery:** This process identifies data sequence patterns based on a series of data.

## **MACHINE LEARNING**

Machine Learning use the process one step ahead than Artificial Intelligence by contributing the information for a machine to learn and adapt for exposing to new dataset. In Machine learning we give the complete information to the machine to train it completely to handle a new situation. It first read the data mined data, and then use a new algorithm created by AI to update current algorithms with the help of learned new task. Machine learning and data mining both use the same algorithms to learn relationships in the data. Machine learning is branch of manmade brainpower science (artificial intelligence) i.e. the frameworks that can learn information.

### **Machine Learning techniques**

Following are the four types of Machine Learning techniques

- a. **Supervised Machine Learning Algorithms:** Supervised Machine learning algorithms are built to analyse known types of data and make predictions on given set of samples. This technique typically builds predictive models based on known data sets and its relevant predictions. This algorithm looks for patterns within the value labels associated to data points.
- b. **Unsupervised Machine Learning Algorithms:** Unsupervised Machine Learning Algorithm is used for unstructured and complex data which doesn't have data labels. These machine learning algorithms systematically categorises data into a group of clusters to define its structure. By using clustering techniques any complex data structure can be examined quickly.
- c. **SemiSupervised Machine Learning Algorithms:** This methodology is typically a combination of supervised and unsupervised learning. This technique used to standardize the data which contains large unlabelled data along with small labelled data. This technique stabilise the complex data to a large extent using both supervised and unsupervised techniques.
- d. **Reinforcement Machine Learning Algorithms:** Reinforcement Machine Learning Algorithms is selfevolving techniques which change strategies based on past actions and related results. These algorithms choose actions based on available data point and then analyses the decision was. Learning from past actions and decisions are used to improve strategy and attain better results.

## **ARTIFICIAL INTELLIGENCE**

AI is a way to make machines think and behave intelligently. These machines are controlled by software inside them, so AI has a lot to do with intelligent software programs that control these machines. It is a science of finding theories and methodologies that can help machines understand the world and accordingly react to situations in the same way that humans do.

Data mining, artificial intelligence and machine learning are so interconnected that it is actually very hard to form a hierarchy between these three. Actually they are involved in interdependent relation so that a mixture of their methods can be used to design precise outcomes. Data mining is always used as an essential measure for coding programs with the information and statistics required by AI to generate a solution.



### **Common types of Artificial Intelligence Techniques**

- a. Case-based reasoning: Case-based reasoning (CBR) solves a problem by recalling similar past problems assumed to have similar solutions. Numerous past cases are needed to adapt their solutions or methods to the new problem. CBR recognises that problems are easier to solve by repeated attempts, accruing learning.
- b. Rule-based expert system: This technique is the simplest form of AI. It has predefined rules and techniques. Each rule has an IF conditional statement, as the IF criteria matches, actions related to respective condition is executed.
- c. Artificial neural networks: It is based on the working of human brain by making the right connections at given condition. Artificial neural networks consist of multiple nodes as biological neurons of human brain. The nodes are connected by links to interact with each other. Every link is having some weight. By changing the link weight ANN can accomplish learning. The nodes can perform simple operations on the input data and can share results to other neurons. The result of node is known as activation or node value. Neural network is another emerging technique used in Educational data mining. ANN learning can be either supervised or unsupervised.
- d. Genetic algorithms: It is based on evolutionary biology techniques such as selection, mutation, inheritance and recombination. Genetic algorithms are used for searching through large and complex data sets for finding optimized solutions. They give solution for search problems based on the theory of natural selection and evolutionary biology.
- e. Fuzzy systems: It is one of the reasoning methods based of human reasoning system. It includes all the values between zero and one where zero represents false and one represent true. In fuzzy system the logic cannot give only absolute truth and absolute false value. But intermediate values are also possible which is partially true and partially false.
- f. Multi-agent systems: It is system which allows independent processes to coordinate with each other. In this system the problem knowledge is divided into the necessary subunits. Then the independent intelligent agents coordinate the agent activities. It supports the feature of distributed artificial intelligence.
- g. Reinforcement learning: Reinforcement learning is often used for robotics, gaming and navigation. With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards.

### **DATA SCIENCE**

Data science involves Machine learning as well as statistics methods to analyse data for making future prediction. It is a practical application of machine learning with an objective of solving real world problems. It is an interdisciplinary area with processes and systems to search knowledge from dataset in various formats. It helps AI to provide solutions to problems by associating similar data for prospect use. Data science is the application of quantitative and qualitative methods to solve relevant problems and predict out-comes.

Common types of data science techniques:

- a. Anomaly Detection: Anomaly Detection searches for information in a set of data, which does not match expected behaviour or predicted usual pattern. Anomalies are also termed as exceptions, contaminants, outliers or surprises and they usually denote actionable important information. These objects could considerably deviate from the usual average of a dataset. Anomaly detection data which does not fits in the data set and shows considerable deviation which suggest that there are unknown facts in the data which needs more attention.
- b. Clustering Analysis: Clustering Analysis classifies data sets with same characteristics and studies their similarities along with the differences with other data sets. It forms cluster of similar data based on their attributes. Each cluster has common traits which can be further used to improve algorithm. As an example we can observe customers buying same type of products or buying on similar dates of the



month. This can boost the promotion techniques to focus on these clusters for specific products or specific shopping period.

c. Association Analysis: Association Analysis discovers relevant associations between different variables in a large database. This technique helps to ascertain patterns in the given data which can detect variables in the data and establish connections between various variables. This technique is typically used in retail stores to look for buying patterns of the customers and accordingly outlet can send recommendation of similar products the customer has bought earlier or recommend supplementary product for the items customer has already bought. This technique is very useful to analyse large data and find similar patterns occurring in regular intervals.

d. Regression Analysis: Regression Analysis determines dependency between attributes. It is assumed that one attribute has a single way casual effect in response to another attributes. Two attributes can be related to each other by one way effect which forms dependency. Independent attributes can affect each other, but this doesn't necessarily denote that both are dependent on each other. There is always a possibility that one attributed is dependent on other but not the same way around. Regression analysis determines this one way effect between the attributes and find out which variable is dependent on another.

#### **Classification Analysis:**

Classification Analysis systematically segregates large data to classify it into different categories. This technique is helpful to identify different types of data present in the data set and among those which are significant data categories which needs further analysis. Classification analysis is normally implemented with cluster analysis as classifying data is usually the pre-requisite for clustering. Email classification can be a good example of these techniques. Using keywords or sender information mails can be classified as internal, classified or spam. By categorising these mails user can easily focus on important mails first to improve efficiency and effectively manage inbox.

#### **Linkage between Data Mining, Machine learning, Artificial Intelligence and Data Science methodologies**

Data mining is a technique of collecting historic data and segregating it into different clusters based on its characteristics. Data mining also ascertain patterns in large data sets based on occurrences and frequencies. Data mining is an effective technique of data classification and to present complex data into a systematic database with data labels. Machine learning is used to predict future events based on historic data patterns. It uses existing data sets to perform analysis to form rules and actions for future events. Machine learning uses supervised, unsupervised learning algorithms to interpret data by applying predefined rules. It also evolves based on learning from past actions and its respective results. Using Reinforcement Machine Learning algorithms, strategies and actions can be improved to optimize the performance. Artificial Intelligence is a broader concept also involving actions based on data segregations and predictions. It can be considered as a self-sufficient independent authority to execute or recommend certain course of actions based on the scenario. Robotics, BOTs are some of the systems based on AI. Data Science on the other side critically examines complete scenario and actions to understand the dynamics of the overall transaction. It evaluates the results and gain insight on further improving the process. Using multiple analytical techniques Data Science shows common attributes which are not yet defined or related attributes which has one way impact on another. It is very important for every process to identify exceptions and outliers which are not in usual average. Outliers are identified with the help of Data Science techniques. These are indications of certain anomaly which needs more attention.

#### **CONCLUSION**

Data mining, Machine learning, Artificial Intelligence and Data Science methodologies have their independent features, while building a system, all these techniques go hand in hand. All these methodologies are interrelated as well as dependent on each other to perform their respective tasks. We can tie these four techniques with common thread of data. Data produced by Data Mining is used



by Machine Learning to predict future event and relevant rules. Artificial Intelligence executes the action based on Machine Learning rules and understanding. Once the operation is completed related actions and results are studied using Data Science techniques to optimise and improve the performance. Data Science covers Artificial Intelligence techniques, which contains machine learning techniques. Machine learning contains deep learning which is based on data mining data.

## REFERENCES

1. M. H. Dunham, Data Mining: Introductory and Advanced Topics, Pearson Education, 2010, ISBN: 0-13- 088892-3
2. Nawal Ali Yassein et. al. (2017), Predicting Student Academic Performance in KSA using Data Mining Techniques, Journal of Information Technology & Software Engineering, ISSN: 2165-7866
3. Siddhesh Chavan et. al. (2018), Data Mining Techniques to Improve Customer Relationships Management, International Research Journal of Engineering and Technology (IRJET), Volume: 05 Issue: 02, e-ISSN: 2395-0056
4. Hari Prasad S et. al.(2016), EVALUATION OF MACHINE LEARNING ALGORITHMS IN ARTIFICIAL INTELLIGENCE, International Journal of Pharmacy & Technology, ISSN: 0975-766X
5. Prateek Joshi (2017) , Artificial Intelligence with Python, ISBN 978-1-78646-439-2
6. S.H. Chen et al. (2008), Artificial Intelligence techniques: An introduction to their use for modelling environmental systems, Mathematics and Computers in Simulation 78 (2008) 379–400
7. Wilton W.T. Fok et. al. (2018), Prediction Model for Students' Future Development by Deep Learning and Tensorflow Artificial Intelligence Engine, 4th IEEE International Conference on Information Management, 978-1-5386-6147-5/18/\$31.00 ©2018 IEEE
8. Samy Abu Naser et. al. (2015), Predicting Student Performance Using Artificial Neural Network: in the Faculty of Engineering and Information Technology, International Journal of Hybrid Information Technology Vol.8, No.2 (2015), pp.221-228, ISSN: 1738-9968 IJHIT
9. Pariwat Ongsulee (2017), Artificial Intelligence, Machine Learning and Deep Learning, Fifteenth International Conference on ICT and Knowledge Engineering, 978-1-5386-2117-2/17/\$31.00 ©2017 IEEE
10. Matthew A. Waller and Stanley E. Fawcett (2013), Data Science, Predictive Analytics, and Big Data: A Revolution That Will Transform Supply Chain Design and Management, Journal of Business Logistics, 2013, 34(2)