ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 1, January 2017

A Study on Basics of Data mining, Machine Learning and Big data

J.Vasuki[1], S.Priyadarshini[2]

Assistant Professor, Dept. of I.T., SNS College of Technology, Coimbatore, Tamilnadu, India

Assistant Professor, Dept. of I.T., SNS College of Technology, Coimbatore, Tamilnadu, India

ABSTRACT: Data mining is a process that turns raw data into useful information. Machine learning algorithms help in analyzing data and finding useful models, pattern and other regularities in data. Data mining and Machine learning go together in almost many areas. Whenever a mining algorithm tries to find a pattern or predict some details it trains the machine to do, this implies that machine learning plays a vital role in mining useful information. Because of the vast increase in data the traditional method of mining is not sufficient. Hence now the concept of data mining has been upgraded to Big data analytics and the researchers are now turning their focus towards the study of big data. This paper focus on the study of basic concepts of these three important domains, since learning the basic provides strong foundation for applying the concepts effectively

KEYWORDS: Data Mining, Machine learning, Big data, Supervised/Unsupervised learning,

I. INTRODUCTION

Data is increasing rapidly day to day due to rapid development of information technology and its usage by the public. Useful information can be obtained when these raw data's are studied properly. Data mining is a process that turns raw data into useful information. The unprocessed data's need to be collected, stored and maintained properly before they are applied to any of the mining techniques. These mining techniques automatically searches large store of data and discover patterns. Data mining depends on effective data collection and warehousing as well as computer processing.

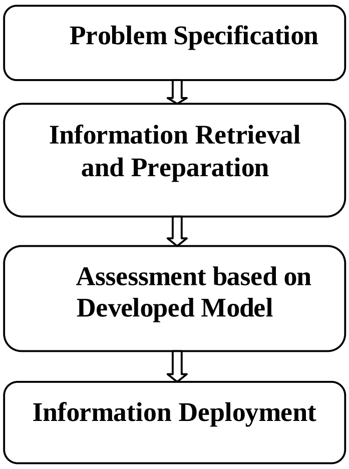
Problem Specification
Information Retrieval and Preparation
Assessment based on Developed Model
Information Deployment

Figure 1 The Data Mining Process

II. LITERATURE REVIEW

Puneet Singh Duggal, Sanchita Paul, “Big Data Analysis : Challenges and Solutions”, international Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV, presents various methods for handling the problems of big data analysis through Map Reduce framework over Hadoop Distributed File System (HDFS). Map Reduce techniques have been studied in this paper which is implemented for Big Data analysis using HDFS.

Chanchal Yadav, Shullang Wang, Manoj Kumar, “Algorithm and Approaches to handle large Data- A Survey”, IJCSN, Vol 2, Issuue 3, 2013 ISSN:2277-5420, presents a review of various algorithms from 1994-2013 necessary for handling big data set. It gives an overview of architecture and algorithms used in large data sets. These algorithms define various structures and methods implemented to handle Big Data and this paper lists various tools that were developed for analyzing them. It also describes about the various security issues, application and trends followed by a large data set [9].

Wei Fan, Albert Bifet, “Mining Big Data: Current Status, and Forecast to the Future”, SIGKDD Explorations, Volume 14, Issue 2, presents a broad overview of the topic big data mining, its current status, controversy, and forecast to the future. This paper also covers various interesting and state-of-the-art topics on Big Data mining.

Priya P. Sharma, Chandrakant P. Navdeti, “Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution”, IJCSIT, Vol 5(2), 2014, 2126-2131, discusses about the big data security at the environment level along with the probing of built in protections. It also presents some security issues that we are dealing with today and propose security solutions and commercially accessible techniques to address the same. The paper also covers all the security solutions to secure the Hadoop ecosystem.

Richa Gupta, Sunny Gupta, Anuradha Singhal, “Big Data : Overview”, IJCTT, Vol 9, Number 5,March 2014.This paper provides an overview on big data, its importance in our live and some technologies to handle big data. This paper also states how Big Data can be applied to self-organizing websites which can be extended to the field of advertising in companies

III. DATA MINING PROCESS DESCRIPTION

The process involved in data mining is given in the figure 1.

A. PROBLEM SPECIFICATION

This initial phase of a data mining deals with focusing on knowing the requirements and specification. Once you have specified the project from an industrial perception, the data mining problem could be devised to develop a preface execution plan.

A problem for example would be how to increase sales of a product. In such cases, data mining comes into consideration. A model would be constructed based on number of people who purchase the product. A survey has to made on the people who have purchased the product and people who have not purchased the product. Attributes of the customers could be age, name, residential address etc.

B. INFORMATION RETRIEVAL AND PREPARATION

This phase involves obtaining and gathering of data to address the problem. At this stage the unwanted data has to be removed and only the needed data is maintained involving additional data. It also performs identification of data rich problems and to locate for patterns in the repository.

In data preparation a sample model has to be built such that it maintains the needed data to create case table. This phase involves designing of the model for multiple times without considering any specific order. Modeling involves table, cases and choosing attribute. Data cleaning and transformation are also involved at this stage. The unwanted data are removed and only the needed data are kept. Example includes replacing monthly income to annual income, initials with first name etc.

New attributes that may be relevant to the data may be included to obtain more information. The number of purchases and comments can be included instead of amount of purchase to get better information. Hence considerate data and information helps a lot for data mining

C. ASSESSMENT BASED ON DEVELOPED MODEL

In this stage various techniques are applied and the parameters are standardized to finest value. Model development at initial stage involves testing with very few data and the final case table could include large set of data. This phase estimates whether the requirement meets the ultimate goal and if exact results are produced. It predicts queries of customers. Is the customer is likely to do product purchase? Is there a need to add extra data (column)? Number of products sold.

D. INFORMATION DEPLOYMENT

Information deployment is making use of data mining within a specific location where the actionable information can be obtained from the data. Deployment can include developed model elicitation, integration of data mining models with various applications, framework of Data Warehouse, or tools for Oracle data mining develops and applies mining inside their database and makes the results available. It also provides real time scores. Example a proprietor could run a model that indicates if a company that he could deals with contracts is fake or genuine

IV. DATA MINING OBJECTIVES

Data mining deals with the set of data that can be modeled and figured out to produce desired results. The data mining objectives are categorized as: Supervised and unsupervised data mining. Supervised and unsupervised learning are natives of Machine Learning which comes under Artificial Intelligence.

Artificial Intelligence is the development and execution of computer systems based on the environment that performs assigned jobs and produces best results. AI makes machines function like humans. And Machine Learning implements devices to learn from their own results and alter their working functionality.

- *Supervised Data Mining*

Supervised learning which is a type of Machine Learning that deals with obtaining a function from obtained labeled data. This analyses the training data and produces a function which maps to produce new desirable results based on previously known attribute. Directed data mining explains the nature of the objective as a part of attributes that are independent.

Supervised learning produces outcomes to predictive models. Construction of supervised model associates in training, the software inspects many options where the final value is known already. The model "learns" the philosophy for producing the prediction in Training phase. Example a model that obtains promotion must be trained by interpreting the aspects of people who have responded or who have not responded previously.

- *Unsupervised Data Mining*

Unsupervised learning is another type of Machine Learning that produces data from functions that involve hidden structure. There is no difference between dependent and independent attributes. No previous results are present to build models for the algorithms used. Unsupervised learning can be used for descriptive purposes and also predictions.

V. DATA MINING TASKS

Data mining which is otherwise called as data or knowledge discovery is the method of evaluating and classifying data from various prospect and encapsulate it into useful information

In short, data mining is the method used to find the correlations or patterns from enormous amount of range in a huge database.

Data mining involves the following methods:

- Classification
- Estimation

- Prediction
- Association rules mining
- Clustering
- Anomoly Detection

DESCRIPTION

A. Classification

Classification is a method that involves deriving and dividing the data bases on various occurrences and existing data. There are several types of classification algorithms that fall under data mining. They are Decision tree, k-nearest neighbor classifier, Naive Bayes, Apriori and AdaBoost. Classification consists of investigating the characteristics of a newly obtainable object and transferring to it a predefined class. Classification task is defined by the precise classes, and a training set involving examples that are classified again.

Examples of classification algorithms include:

- Linear classifiers
  - Fisher's linear discriminant
  - Logistic regression
  - Naive Bayes classifier
  - Perceptron
- Support vector machines
  - Least squares support vector machines
- Quadratic classifiers
- Kernel estimation
  - k-nearest neighbor
- Boosting (meta-algorithm)
- Decision trees
  - Random forests
- Neural networks
- FMM Neural Networks
- Learning vector quantization

B. ESTIMATION

Estimation is a technique of dealing with constantly valued results. Based on the given input data, estimation is used to obtain values for unknown variables. Examples are income, age, height etc

C. Prediction

It is the method that which describes how and what way the things will happen, based on knowledge or involvement. Prediction may be a statement in which some result is expected. It may be used to find unknown or missing values. Examples are medical diagnosis, fraud detection.

D. ASSOCIATION RULE MINING

An association rule is a method which implies association relationships among a set of objects or variables in a database. Association rule mining analyses data based on support and confidence. Frequency of item in database is support and if the product is genuine it is confidence. Some of the association rule mining algorithms widely used are:

- Apriori
- Partition
- FP growth
- ECLAT

E. CLUSTERING

Clustering is the most essential method of unsupervised learning that focuses on group formation based on certain features like characteristics and similarities. It deals with finding or forming of a structured from a set of data that are not labeled. Clustering can be centralized, decentralized, graph, group and so on. It is mainly used in robotics, mathematical analysis etc.

Algortihms used for clustering are

- Artificial neural network (ANN)
- Nearest neighbor search
- Neighborhood components analysis
- Latent class analysis
- K-Means

F. ANOMOLY DETECTION:

Anomaly detection is also called as outlier detection. This method deals with finding of items, interpretation or events that does not match an estimated prototype or item in the dataset. The major categories of Anomaly detection are: unsupervised Anomaly detection, Semi- Supervised Anomaly detection, and Supervised Anomaly detection. It is mainly used in health monitoring, Fraud detection etc.

The most popular techniques of Anomaly detection are:

- Density-based techniques (k-nearest neighbor, local outlier factor, and many more variations of this concept).
- Subspace- and correlation-based outlier detection for high-dimensional data.
- One class support vector machines.
- Replicator neural networks.
- Cluster analysis-based outlier detection.
- Deviations from association rules and frequent item sets.
- Fuzzy logic based outlier detection.
- Ensemble techniques, using feature bagging, score normalization and different sources of diversity.

The categorizations of the data mining algorithms are depicted in the table below

TASK	**ALGORITHM**
Association	Apriori, Partition, FP growth, ECLAT
Clustering	K-Means, Expectation Maximization, DB SCAN, fuzzy C Means.
Classification	Decision Trees, C4.5, KNN, Naïve Bayes, SVM.
Regression	Multivariate Linear regression

Table1- Classification of algorithm

VI. MACHINE LEARNING

Machine Learning is a sub-field of data science that focuses on designing algorithms that can learn from and make predictions on the data. There are numerous applications of machine learning. It is hard to understand how much the machine learning has achieved in real world applications. Machine learning is generally just a way of fine tuning a system with tunable parameters of the data. It is a way to make the system better with examples, usually in a supervised or unsupervised manner.

Types of Machine learning

- Supervised
- Unsupervised
- Regressive

MACHINE LEARNING PROCESS

Step 1: Data Initialization

In this step, the data on which the algorithm to be applied is to be collected and organized. The collected data need to be preprocessing in order to remove noise. Feature selection is to be made for getting effective output. For this data mining algorithms can be used.

Step 2: Learning Model

Machine learning models will be prepared with certain combination of algorithms and the training data sets

Step 3: Testing Model

The finalized machine learning model will be tested with the test data sets to check the efficiency. Step 2 will be called for when step 3 fails to give the expected output. The process will be repeated until the final model is obtained

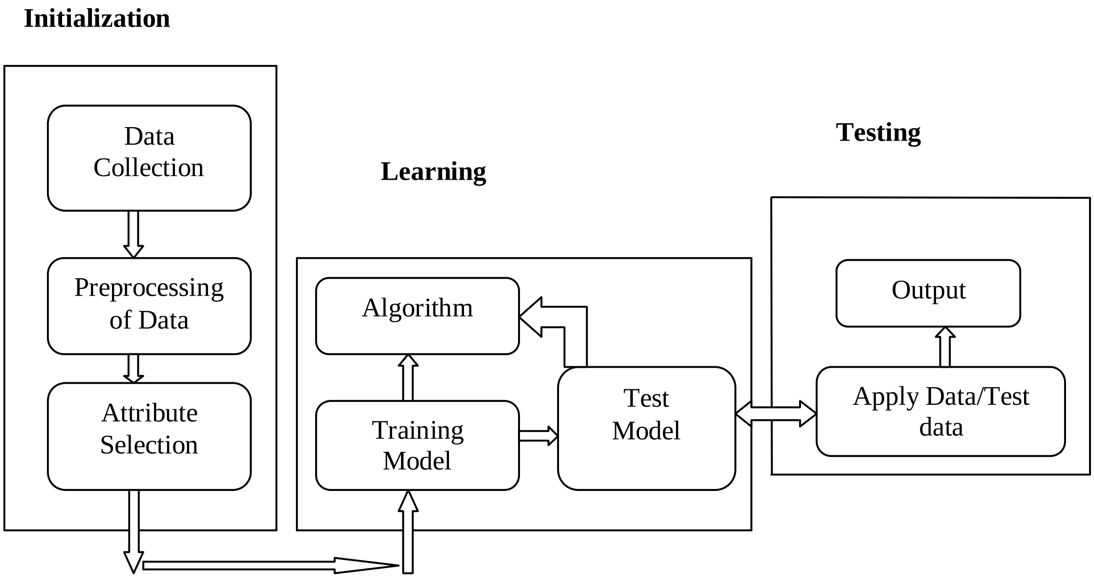

Figure 2. Machine learning Process

Machine learning is applied in the offline training phase and is used to improve the following applications.

1. Face detection: Cameras can automatically snap a photo when someone smiles this is because of the face detection feature in mobile cameras which is an example of machine learning.
2. Face recognition: Machine learning algorithm trains the system to identify an individual from a photo this feature is used in social sites for automatically tagging people in photos where they appear.
3. Image classification: Machine learning helps to improve image classification or image categorization in large sets of images.
4. Speech recognition: Improvements in speech recognition systems has been made by, machine learning
5. Anti-virus: Machine learning is used in Anti-virus softwares to improve detection of malicious software on computer devices.
6. Anti-spam: Machine learning is also used to train better anti-spam software systems.
7. Genetics: Classical data mining or clustering algorithms in machine learning such as agglomerative clustering algorithms are used in genetics to help find genes associated with a particular disease.
8. Signal denoising: Machine learning algorithms such as the K-SVD which is just a generalization of k-means clustering are used to find a dictionary of vectors that can be sparsely linearly combined to approximate any given input signal. Thus such a technique is used in video compression and denoising.
9. Weather forecast: Machine learning is applied in weather forecasting software to improve the quality of the forecast.

Machine learning helps develop sophisticated software systems without much effort on the human side. Instead of spending years handcrafting features or fine tuning a system with a lot of parameters, machine learning does that quicker. It also only requires training data to learn better features or parameters needed to improve a given system.
The only drawback in it is that machine learning doesn't work well for non-convex problems or problems with discrete parameters that are not differentiable.

MACHINE LEARNING MODELS:

There are different options to Machine Learning Models based on classification and regression. Each model is viewed as black-box to solve same problem. But each and every model comes from different algorithms and produce results based on various methodologies used. However, each model come from a different algorithm approaches and will perform differently under different data set. An analysis can be made to find which algorithm is best on test data.
Approaches of Machine Learning include:

- Artificial neural networks
- Decision tree learning
- Association rule learning
- Deep learning
- Inductive logic programming
- Support vector machines
- Clustering
- Bayesian networks
- Reinforcement learning
- Representation learning
- Genetic algorithms
- Rule-based machine learning
- Learning classifier systems

VII. BIG DATA HISTORY AND EVOLUTION

Big data analytics examines huge quantity of data to discover hidden patterns, correlations and other insights. With improvement in technology data analysis and information retrieval has become an efficient task.
Organizations now focus on large amount of data since it help them apply analytics and obtain signification results from it. But in 1950's big data was all about the usage of more number of worksheets done manually. But after the improvement of technology and data, information analysis, maintenance, decisions and obtaining of expected outcome has become easier leading to stronger and competitive edge.
Why is big data analytics important?

Big data analytics provides methods for organizations control their data and use it to discover new opportunities. This leads to smarter business operations, increased profits, highly proficient operations, and satisfactory customers. IIA Director of Research Tom Davenport understood the use of big data after having to know more number of industrial methods.

1. Cost reduction. Big data techniques deal with large number advantages when it comes to cost and storage.
2. Faster, improved decision making. The speed of Hadoop and in-memory analytics helps to analyze new sources of data and make improved and efficient decisions.
3. New products and services. Based on customer needs and satisfaction new products can be introduced leading to increase in productivity and service.

VIII. CONCLUSION

Due to the development of technology, data sharing, and social sites the accumulation of data is increasing tremendously day by day. Managing and handling of these enormous data is a challenging job. In this paper we discussed about the various process and methods of data miming and how it is related to Machine Learning. Hence data are needed for making any machine to work and produce results. Though data mining and machine learning are not same they are similar and work with each other. Machine learning can be used for data mining and data mining makes use of methods that operate on top of Machine Learning. Big data plays vital role in handling large data sets to take decisions, classifying, predicting, modeling, and finding new and expected outcomes. Big data is used in wide areas of scientific research to find frequent patterns and rules that are hidden. Big data analysis helps industries to take improved decisions, to predict and discover changes and to recognize new opportunities.

REFERENCES

1. Bharti Thakur, Manish Mann, “ Data Mining for Big Data: A Review”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 5, May 2014
2. Pang-Ning T, Steinbach M, Kumar V.,” Introduction to data mining”, 2006.
3. Xindong Wu, Xingquan Zhu, Gong Qing Wu, Wei Ding,”Data mining with Big data‟, IEEE, Volume 26, Issue 1,January 2014.
4. Bharti Thakur, Manish Mann,”Data Mining for Big Data- A Review‟, IJARCSSE, Volume 4, Issue 5,May 2014.
5. Rohit Pitre, Vijay Kolekar, ‟A Survey Paper on Data Mining With Big Data‟, IJIRAE, Volume 1, Issue 1, April 2014.
6. Dr. A.N. Nandhakumar, Nandita Yambem, “A Survey of Data Mining Algorithms on Apache Hadoop Platforms‟, IJETAC,Volume 4, Issue 1, January 2014.
7. C.L. Philip Chen, C.-Y. Zhang, “Data-intensive applications, challenges, techniques and technologies: A survey on Big Data”,Inform. Sci. (2014), http://dx.doi.org/10.1016/j.ins.2014.01.015.
8. Alex Berson and Stephen J.Smith,” Data Warehousing,Data Mining and OLAP”, edition 2010.
9. Chris Bozzuto. "Machine Learning: Genetic Programming.", February 2002
10. Hsinchun Chen. "Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms" ,available at http://ai.bpa.arizona.edu/papers/mlir93/mlir93.html#318.
11. Yisehac Yohannes, John Hoddinott " Classification and Regression Trees- An Introduction" International Food Policy Research Institute, 1999.
12. Xindong Wu , Gong-Quing Wu and Wei Ding “ Data Mining with Big data “, IEEE Transactions on Knoweledge and Data Enginnering Vol 26 No1 Jan 2014
13. Puneet Singh Duggal, Sanchita Paul, (2013), “Big Data Analysis:Challenges and Solutions”, Int.Conf. on Cloud, Big Data and Trust, RGPV
14. Chanchal Yadav, Shullang Wang, Manoj Kumar, (2013) “Algorithm and Approaches to handle large Data- A Survey”, IJCSN, 2(3), ISSN:2277-5420(online), pp2277-5420
15. Richa Gupta, Sunny Gupta, Anuradha Singhal, (2014), “Big Data:Overview”, IJCTT,
16. Richa Gupta, (2014), “Journey from data mining to Web Mining to Big Data”, IJCTT, 10(1),pp18-20