AN ANALYSIS OF BIG DATA ALGORITHMS (SUPERVISED, UNSUPERVISED & REINFORCEMENT): A SURVEY

Mr.K.Jamberi [a,*], Mrs.S.Kalaiselvi [b], Mrs. B. Sita Devi [c], Mrs.R.Shyamala [d]
[a,b,d]Assistant Professor , K.C.S.Kasi Nadar College of Arts & Science ,
[c] Ph.D. Research Scholar,VEL'S University.

ABSTRACT

Big data has gained popularity in the last decade due to increased demand to process large data sets to deal with ever-growing volumes of data and the extraction of key information from massive volumes of data. Big data technologies are being widely found in many application domains that produce large levels of data that demand efficient data processing algorithms which co-relates different fields. With the complex characteristics of big data, new problems have emerged and need to face new challenges when developing and designing a brand-new algorithm for Big Data Analytics. Currently, there exist several algorithms which differ by their application area and efficiency. Increase in computational power and algorithmic improvements have reduced the full time for big data sets. The key benefit of the algorithm is that it makes possible to process the elements of large data sets in big data, extract the value from big data, handling any size of big data set with assistance from algorithms. This paper proposes algorithms their processing speed, efficiency and accuracy. It is important to review the algorithms to create big data valuable.

Keywords: Big Data; Algorithms; Machine Learning.

1. INTRODUCTION

Big data is a term for large data sets have a larger, more varied and complex structure with difficulty storing, analysing and visualizing for further processing or result. Big data required technologies to capture, storage, management, and analysis the huge amount of data with more complexity and it can be characterized by its size and variety [1]. It is a term that describes a large volume of data - structured data along with unstructured and semi- structured data. Data are needed to isolate the hidden patterns and to find answers without over-fitting the data. The research process has large amounts of data to reveal hidden patterns and correlations secret named as big data analysis. These data sets are so voluminous that traditional data processing software just can't manage it. But these large volumes of data can be used to address the business problem. Systems that process and store large amounts of data has become a common component of the architecture of data management within the organization. Big data often characterized by 3Vs: large volumes of data in multiple environments, various types of data stored in large data systems and the speed at which data is generated, collected and processed. The sharp increase flood of data in the big data era brought great challenges in data acquisition, storage, management and analysis. The evolution of big data driven by the rapid growth in demand for cloud computing applications and development of virtualization technology. Therefore, cloud computing not just provides computation and processing for big data, but additionally itself is just a service mode. To a specific extent, the advances of cloud computing also promote the development of big data, both that supplement each other. At present, the info processing capacity of IoT has fallen behind the collected data and it is extremely urgent to accelerate the introduction of big data technologies to advertise the development of IoT. Keeping up with big data technology and algorithms to process very large data is an ongoing challenge.

Machine Learning aims to develop programs that allow machines to learn and make decisions without human intervention to learn and grow, machines need data to analyse and to train on which explains the renewed interest of Machine Learning with the arrival of Big Data, Learning methods for Machine Learning can be categorized as supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning. One of the main tasks so supervised learning is the classification. While, the clustering presents the most important technique among those of unsupervised learning [2]. Algorithms take a combination of both - the input and output simultaneously to "learn" the data and output results when given a new input. Big data processing using Machine learning allows a computer system to make predictions or take some decisions using historical data without explicitly programmed. It uses large amounts of structured data and semi-structured so that machine learning models can produce accurate results, or to make predictions based on data. In Supervised Learning, as the name rightly suggests, it involves making the learning algorithm of data while providing the correct answer or the label for data. This means that a class or a value that would be predicted well known and well defined for the algorithm from scratch. Another class is under Unsupervised Learning, where, unlike the method of supervised algorithms that do not have a correct answer or no answer at all, it's up to the wisdom of an algorithm to combine the same data and understand it. Along with these two prominent classes, we also have a third class, called Reinforcement Learning. In Reinforcement Learning, there are rewards given to the algorithm on any precise predictions thus encouraging a higher accuracy. While the above three classes covering most fields in a comprehensive manner, we sometimes still have the problem of land into a lump of performance of our model. In such cases it may make sense to use the ensemble method to obtain higher accuracy. The algorithm in each category, in effect, doing the same task to predict the output given input is not known. Ensemble learning uses multiple models to obtain better predictive performance than could be obtained from any of the constituent models presents four methods of combining multiple models: bagging, boosting, stacking and error-correcting output [3]. the critical issues of machine learning methods for big data from five different perspectives, including learning for large scale of data, learning for different types of data, learning for top speed of streaming data, learning for uncertain and incomplete data, and learning for extracting valuable information from massive amounts of data however, here the data is the key driver when it comes to choosing the right algorithm that fall under the major three categories like classification, regression and clustering problems.

Many applications produce large amounts of data that demand efficient data processing algorithms. Currently, the design and implementation of Machine Learning algorithms has become a tedious task especially in this era of Big Data characterized by the enormous volume of data, their high speed of production and diffusion as well as their very varied nature, therefore, to face these new challenges, it is essential to look for new algorithms suitable for Big Data or to review the classic algorithms to adapt them to this new context in order to, effectively, manage and analyse big data [2]. Traditional machine learning algorithms are mostly based on memory, but not all the Big Data is stored in memory, there is a need to propose new algorithms to meet the needs of Big Data processing. In the research of Big Data correlation analysis, it mainly comes from Apriori algorithm in the Big Data classification method, SVM learning method, which is fast, stable and robust. a decision tree from large-scale data, which can train the samples without saving the training set data in the memory and improve the operation speed. A K-means algorithm based on MapReduce, which has better performance in processing speed and processing scale also, based on MapReduce framework, a parallel KNN algorithm, which is efficient and scalable in large-scale data processing [4]. In the present investigation an attempt has been made to give review of algorithms which is very useful in all the field to analyse, store, and manage the big data.

2. LITERATURE REVIEW

Big Data increasingly benefits both research and industrial areas such as health care, finance service and commercial recommendation. The Economist says, Data are becoming a new raw material of business. Economic information is practically identical to capital and work. Nowadays, the data to be analyzed are dynamic and huge in volume, also they are the group of different data types. These data come from different data sources such as WhatsApp, Twitter, Facebook, YouTube, Mobile phones GPS signals and more. Hence, the Big Data has the unique features such as heterogeneous, unstructured, semi structured, incompleteness, high dimensional. Big Data has complex natures that need powerful technologies and advanced techniques. So, the traditional static Business Intelligence tools can no longer be resourceful in the case of Big Data applications.[5]

A systematic literature review as the process of identifying, assessing and interpreting all research results to provide answers to research question consists of several activities, namely: specifying the research questions, selecting studies, extracting required data, synthesizing data and describing the result. [6]

This article reviews the state of the art for the application of Big Data. Initially, an overview of Big Data and its features are presented. Then, the main aspects of Big Data processes and technologies are discussed. Afterwards, relevant applications of Big Data analytics are discussed. Next, Big Data Analytics are discussed in general terms and especially for the healthcare sector. The article ends with a review of the challenges that were identified in this study, followed by the conclusions.[7]

Harshawardhan S. Bhosale, Devender P. Gadekar explained briefly Big Data and 3V's.This paper explained various problems facing with Big Data processing like heterogeneity and incompleteness of data, Scale, time of analysis and security and Privacy of data etc. They explained Hadoop as a solution of Big Data processing. This paper explained Hadoop architecture as well as HDFS and Map reduce architecture in brief. At the end of the paper they explained various components on the basis of concurrency, durability, replication methods, database model and concurrency etc [8].

Cheikh Kacfah Emani, Nadine Cullot, Christopher, Nicolle et.al reviewed the idea of Big Data.They discussed the features of Big Data and also explained the steps of Big Data processing. During the management of Big Data many problems can be encountered during semantic gathering. They also explained how to tackle Volume, Velocity and Variety [9].

S. Vikram Phannendra,E. Madhusudhana Reddy explained that RDBMS did not provide complete solution while dealing with Big Data. In this paper they described Big Data is different from traditional data in terms of five dimensions this paper also briefly explained Hadoop architecture. Hadoop consists of mainly Name Node, Data Node and Edge Node. Hadoop architecture can handle large dataset, scalable algorithm, log management, Extract –Transform-Load (ETL) platform. They also focused on various challenges of Big Data: Data Privacy, analysis and visualization etc [10].

Vibha Shukla, Pawan Kumar Dubey et.al discussed that data is increasing very fast. With the increased amount of data, Traditional data analysis tools need to develop. In this paper they explained traditional data analytics versus Big Data analytics. They also discussed various Big Data Emerging techniques and technology such as NOSQL database, Map Reduce, Hadoop, HDFS and many more. They also explained various challenges of Big Data and future research will concentrate to develop solutions to deal with these challenges [11].

Prity Vijay and Brigh Keshwani explained the concept of Big Data and briefly discussed problem and challenges faced during the processing of Big Data. Traditional processing and management tools and frameworks are not suitable to handle Big Data because it requires the framework that can handle unstructured data and provide real time analysis with fault tolerance capacity. This paper also explained various problems when we deal with Big Data with traditional approach like RDBMS. Hadoop and its related technologies are suitable for handling Big Data. This paper also explained the various changes that the Hadoop demanded with the time [12].

Raymond Gardiner Goss told about our current position and the way in which we are heading which will increase the need of Big Data management in a faster and secure way. Since, nowadays analysis of more complex data is required, traditional tools are being stressed and moderated. This complex data consists of text, message, photos to information useful for search engines and marketing analysis. Thus, the data is quite unstructured and challenging. Now solutions need to come to tackle the new level of information processing [13].

Poonam S. Patil and Rajesh N. Phursule et.al explained that the major challenge of these days that is the analysis of Big Data. The solution of this challenge is Hadoop framework. Map Reduce programming model is easy to use in parallel and distributed environment. This paper briefly introduced Big Data and Map Reduce framework. Traditional database system does not provide scalable solution that is the requirement of Big Data. This paper also introduced briefly various Hadoop Components [14].

Kuchipudi Sravanthi, T. Subba Reddy explained that Big Data is used in various organizations, companies, enterprises and business etc. They explained various applications of Big Data in various fields. They also explained the usage of Big Data in related fields. At the end they conclude that Big Data is help to makes things easy in various fields that was otherwise very difficult [15].

However, nowadays, many researchers believe that most of this data is unstructured and the use of non-relational (NoSQL) databases is necessary for its management. This data can be categorized into web social media data, machine-to-machine data, transaction data, biometric data and human-generated data.[7]

• **Social media** data is acquired from across interactions, tweets and posts in social networks such as Facebook, LinkedIn, YouTube, and Twitter; • **Machine-to-Machine** data is extracted from machine sensors, meters and other devices;

• **Transaction data** is recovered from fingerprints, genetics, and handwriting and medical images;

• **Human-Generated data** is selected from prescriptions, emails, messages, documents and Electronic Medical Reports;

• **Web data** contains click stream data generated by internet browsers.

3. MACHINE LEARNING ALGORITHMS

A supervised machine learning Algorithm can deal with the obscure ward/target variable which is anticipated from a given arrangement of known indicators (free factors). Additionally, an unsupervised machine learning algorithm can be taken a shot at the objective variable which is anticipated with set of comparable gathering of information things called grouping. Semi-supervised machine learning calculation in which the mix of both Supervised and Unsupervised assortment of information things. Reinforcement machine learning algorithms are given underneath.[21]

4. SUPERVISED MACHINE LEARNING

At the point when you have past information with results (marks in machine learning phrasing) and you need to anticipate the results for the future – you would utilize Supervised Machine Learning algorithms. Supervised machine learning issues can be ordered into 2 sorts - Classification Problems and Regression Problems.

Classification Problems

When you desire to classify results into exceptional classes. For example – whether or not the ground needs cleaning/mopping is a classification problem. The effect can fall into one of the lessons – Yes or No. Similarly, whether or not a consumer would default on their mortgage or now not is a classification problem which is of excessive activity to any Bank.

Regression Problems

Regression additionally a supervised learning calculation for example, order [1]. It finds the connection between some autonomous (known) factors with some reliant (obscure) factors. Instead of downright yield esteem of order, regression gives nonstop quantitative yield esteem dependent on definitely realized information esteems. Some of the capacities, for example, quadratic, cubic, power, logarithmic,

furthermore, others are useful to locate the best estimate of the regression esteem. The Popular Regression Algorithms are successively called Linear Regression and Polynomial Regression.

4.1 LINEAR REGRESSION ALGORITHMS

It finds the relationship between an independent (predictor (X)) and a structured (criterion (Y)) variable to predict the future values of the established variable. Simple regression makes use of one impartial variable and a couple of regressions use two or extra impartial variables to predict the future. Dependent variable has a non-stop and independent variable has discrete or non-stop values. There are two sorts of regression models. One is linear and other one is nonlinear. The linear regression mannequin uses straight line and nonlinear regression mannequin makes use of curved line relationships between based and independent variables.

Place the information of structured and impartial variable values of a precise length in a scatter plot. Correlation with correlation coefficient index values between -1 to +1 of scatter plot will a lot extra useful to visually identify the relationship between the variables. Correlation coefficient indices symbolize the power (strongest) of the relationship between the variables. The easiest correlation coefficient fee of regression between of .5 to .99 is the best prediction indicator of future structured variables. A best t or an fine regression line is wanted to generate in the scatter plot to predict the future variable from the past data.[19]

The Line of Best Fit: A line which has the smallest distance from the information factors in the scatter plot is called regression line. The regression line generally passes through on the imply of structured and independent variables. More distance facts factors are known as error terms. In easy regression these error phrases can also be reachable for the actual world data. A best fitting regression line is drawn with the assist of least-squares method.

Least-squares method: Sum of squares of the difference between every statistics factor to the line. It is additionally called squared error. Regression line may additionally have least squared error. The high-quality t regression line can be prolonged past the historical information place to predict the future facts value. Some independent information values on the scatter plot may additionally be a long way away from the satisfactory t regression line. These statistics factors are called outliers. Sometimes the facts factors may additionally be in a ways away from all the information factors in the horizontal line with less scope. That is additionally referred to as in Outlier observation/outliers. Outliers want to be omitted.[20]

Residuals: Residuals will assist to pick out whether or not there is a linear relationships exist amongst the statistics or not.

Extrapolation: The records factors may additionally be accessible in outside of our taken problem. Example younger infant and the weight calculation trouble may also have elder facts points. These data points are referred to as extrapolation. This will want to be omitted

Equation of a Regression Line:

F (x) = mx + b + e

Where, x = Independent variable

F(x) = Dependent variable

b = y-intercept

m = Slope of the line

e = Error term

Ordinary Least Square (OLS) is useful to minimize the error

(e) cost as lots feasible the usage of the formulation

Σ [Actual (y) – Predicted (y)]2

Preparing Data for Linear Regression

linear regression is been studied at superb length, and there is a lot of literature on how your information ought to be structured to make exceptional use of the model.

Try special preparations of your records the usage of these heuristics and see what works great for your problem.

- **Linear Assumption**: Linear regression assumes that the relationship between you enter and output is linear. It does no longer guide something else. This may additionally be obvious; however, it is appropriate to understand when you have a lot of attributes. You may additionally want to seriously change information to make the relationship linear (e.g. log seriously change for an exponential relationship).
- **Remove Noise:** Linear regression assumes that you enter and output variables are no longer noisy. Consider the use of statistics cleansing operations that let you higher expose and make clear the sign in your data. This is most necessary for the output variable and you prefer to get rid of outliers in the output variable (y) if possible.
- **Remove Collinearity:** Linear regression will over-fit your statistics when you have relatively correlated enter variables. Consider calculating pairwise correlations for you enter facts and getting rid of the most correlated.
- **Gaussian Distributions**. Linear regression will make greater dependable predictions if you enter and output variables have a Gaussian distribution. You can also get some gain the use of transforms (e.g. log or BoxCox) on your variables to make their distribution extra Gaussian looking.
- **Rescale Inputs:** Linear regression will frequently make greater dependable predictions if you rescale enter variables the usage of standardization or normalization.

Advantages

- Shows linear relationship between established and independent variables with superior results.
- A easy mannequin and convenient to understand.

Disadvantages

- It can predict solely numeric output.
- Not relevant for nonlinear data.
- Very plenty touchy with outliers.
- Data should be independent.

4.2. LOGISTIC REGRESSION ALGORITHMS

Logistic regression is used to locate the likelihood of event=Success and event=Failure. We have to use logistic regression when the based variable is binary (0/ 1, True/ False, Yes/ No) in nature. Here the cost of Y tiers from zero to 1 and it can represent through following equation.

odds= p/ (1-p) = chance of match prevalence / likelihood of no longer tournament occurrence

ln(odds) = ln(p/(1-p))

logit(p) = ln(p/(1-p)) = b0+b1X1+b2X2+b3X3....+bkXk

Above, p is the likelihood of presence of the attribute of interest.

Since we are working right here with a binomial distribution (dependent variable), we want to pick a hyperlink characteristic which is satisfactory applicable for this distribution. And, it is logit function. In the equation above, the parameters are chosen to maximize the possibility of gazing the pattern values alternatively than minimizing the sum of squared mistakes (like in normal regression).

Logistic regression is extensively used for classification problems. Logistic regression doesn't require linear relationship between established and impartial variables. It can manage quite a number of kinds of relationships due to the fact it applies a non-linear log transformation to the envisioned odds ratio.

To keep away from over becoming and underneath fitting, we ought to encompass all vast variables. A precise method to make certain this exercise is to use a step smart technique to estimate the logistic regression

It requires giant pattern sizes due to the fact most probability estimates are much less effective at low pattern sizes than regular least square.

The unbiased variables ought to no longer be correlated with every different i.e. no multi collinearity. However, we have the alternatives to encompass interplay outcomes of specific variables in the evaluation and in the model.

If the values of established variable is ordinal, then it is referred to as as Ordinal logistic regression. If established variable is multi category then it is recognized as Multinomial Logistic regression.

4.3. DECISION TREE ALGORITHM

A decision tree is a flowchart-like shape in which every inner node represents a check on a characteristic (e.g. whether or not a coin flip comes up heads or tails) , every leaf node represents a classification label (decision taken after computing all features) and branches symbolize conjunctions of aspects that lead to these classification labels. The paths from root to leaf characterize classification rules.

Decision tree are developed by way of an algorithmic method that identifies methods to break up a fact set based totally on extraordinary conditions. It is one of the most extensively used and sensible strategies for supervised learning. Decision Trees are a non-parametric supervised learning method used for each classification and regression tasks.[18]

Tree fashions the place the goal variable can take a discrete set of values are referred to as classification trees. Decision tree the place the goal variable can take non-stop values (typically actual numbers) are referred to as regression trees. Classification And Regression Tree (CART) is typical time period for this.

Data Format

Data comes in archives of forms.

(x,Y)=(x1,x2,x3,....,xk,Y)

The established variable, Y, is the goal variable that we are attempting to understand, classify or generalize. The vector x is composed of the features, x1, x2, x3 etc., that are used for that task.

Gini Impurity: Gini Impurity is a dimension of the possibility of a wrong classification of a new occasion of a random variable, if that new occasion has been randomly categorized in accordance to the distribution of classification labels from the information set.

If our dataset is Pure then possibility of fallacious classification is zero if our pattern is combination of exceptional lessons then probability of fallacious classification will be high.

Steps for making tree

- Get listing of rows (dataset) which are taken into consideration for making choice tree (recursively at every nodes).
- Calculate uncertainty of our dataset or Gini impurity or how an awful lot our records is combined up etc.
- Generate listing of all query which desires to be requested at that node.
- Partition rows into True rows and false rows primarily based on every query asked.
- Calculate records achieve primarily based on gini impurity and partition of information from preceding step.
- Update easiest records acquire based totally on every query asked.
- Update first-rate query based totally on data achieve (higher statistics gain).
- Divide the node on nice question. Repeat once more from step 1 once more till we get pure node (leaf nodes).

Advantages

- Easy to use and understand.
- Can cope with each express and numerical data.
- Resistant to outliers, subsequently require little records preprocessing.

Disadvantages

- Prone to overfitting.
- Require some sort of dimension as to how properly they are doing.
- Need to be cautious with parameter tuning.
- Can create biased realized timber if some instructions dominate.

4.4 K-NEAREST NEIGHBOR ALGORTHIM (K-NN)

K-NN is a supervised classifier. It is a excellent choice for the classification sort of problems. To predict the target label of a new check data, KNN finds the distance of nearest training statistics classification labels with a new take a look at records factor in the presence of K value. Then ounts the range of very closest statistics factors the use of K price and concludes the new test records classification label. To calculate the range of nearest training information factors distance, KNN makes use of K variable value between zero to 1o normally. Among the famous distance functions like Euclidean distance, Manhattan distance, Minkowski distance and Hamming distance, Euclidean distance characteristic is used for non-stop variables and Hamming distance feature is used for categorical variables.

Let consider, coaching information pattern with n counts. Every statistics factor xi has related category label ci. Here, x denotes coaching records factors and c denotes type labels. For understanding reason plot the education information and its associated classification in (x, y) graph. Also, vicinity the new check data point in the identical (x, y) plan chooses to predict the class label. Now discover the distance between take a look at records factor with all the education facts factors the usage of any one of the distance functions cited in the objective. Arrange instance values in descending order. Now the usage of K variable cost to count the variety of coaching facts factors are close to to test data point. The category label of the most coaching data point inside ok fee will be assigned to classification label of new test data.

Choosing the K value: The challenging phase of the KNN algorithm is to pick the K value. The small K fee in Hence to noise in predicting the goal category label also biggest K price leads to over becoming probability. As properly as biggest K cost will increase the calculation time and reduces the execution speed. So K=n^(1/2) formulation is used to choose the K value. To optimize the check end result cross validation of records will operate on coaching information with different K values. Optimized price will be chosen based on first-rate accuracy for check result.

Condensed Nearest Neighbor: It is the manner of removing undesirable records from coaching statistics to enlarge the accuracy. The steps for condensing the data

- Outliers: Removes the extraordinary distance data.
- Prototypes: To and the non-outlier points, minimum education set is used.
- Absorbed points: Used to perceive non-outlier points correctly.

Advantages

- It is easy to implement.
- It is strong to the noisy coaching data
- It can be greater advantageous if the coaching information is large.

Disadvantages

- Always wishes to decide the fee of K which may additionally be complicated some time.
- The computation price is excessive due to the fact of calculating the distance between the information factors for all the coaching samples.

Random Forest

An enormous number of generally uncorrelated models (trees) working as an advisory group will beat any of the individual constituent models. The low relationship between models is the key. Much the same as how speculations with low relationships (like stocks and securities) meet up to shape a portfolio that is more noteworthy than the total of its parts, uncorrelated models can create group expectations that are more exact than any of the individual forecasts. The explanation behind this brilliant impact is that the trees shield each other from their individual blunders (as long as they don't continually all fail a similar way). While a few trees might not be right, numerous different trees will be correct, so as a gathering the trees can move in the right course. So the essentials for arbitrary woodland to perform well are:

There should be some real sign in our highlights with the goal that models constructed utilizing those highlights show improvement over irregular speculating. The expectations (and in this manner the blunders) made by the individual trees need to have low connections with one another.

4.5 SUPPORT VECTOR MACHINE ALGORITHM (SVM)

Support vector machine is extraordinarily desired through many as it produces large accuracy with much less computation power. Support Vector Machine, abbreviated as SVM can be used for each regression and classification tasks. An SVM mannequin is basically a illustration of special training in a hyperplane in multidimensional space. The hyperplane will be generated in an iterative manner by using SVM so that the error can be minimized. The intention of SVM is to divide the datasets into training to locate a most marginal hyperplane (MMH).

The followings are necessary ideas in SVM −

- Support Vectors − Datapoints that are closest to the hyperplane is referred to as assist vectors. Separating line will be described with the assist of these records points.
- Hyperplane − As we can see in the above diagram, it is a choice airplane or area which is divided between a set of objects having one of a kind class.
- Margin − It may also be described as the hole between two strains on the closet fact's factors of distinctive classes. It can be calculated as the perpendicular distance from the line to the guide vectors. Large margin is viewed as a correct margin and small margin is regarded as a awful margin.

The fundamental purpose of SVM is to divide the datasets into training to locate a most marginal hyperplane (MMH) and it can be carried out in the following two steps −

- First, SVM will generate hyperplanes iteratively that segregates the lessons in satisfactory way.
- Then, it will pick the hyperplane that separates the training correctly.

Advantages

- SVM classifiers provides superb accuracy and work nicely with excessive dimensional space.
- SVM classifiers essentially use a subset of education factors for this reason in end result makes use of very much less memory.

Disadvantages

- They have excessive coaching time therefore in exercise no longer appropriate for massive datasets.
- Another drawback is that SVM classifiers do no longer work properly with overlapping classes.

4.6 NAIVE BAYES ALGORITHM

The naive Bayes algorithm performs classification tasks in the subject of machine learning. It can do classification very nicely on the dataset even it has huge records with multi type and binary category classification problems. The fundamental software of Naive Bayes is text analysis and Natural Language Processing.

Understanding of Bayes theorem will assist to understand (work with) Naive Bayes algorithm efficiently. Bayes theorem is used to mix the a couple of classification algorithms to structure Naive Bayes classifier with a common principle. Bayes theorem works primarily based on conditional probability. Conditional chance means, an event will occur with conditioned (based) on an event already occurred.

Formula:

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

Where,

P(A) = Prior probability of an event A. Here, A is not dependent on an event B in anyway. P(A|B) is a conditional probability of an event A with conditioned an tournament B. If an event A desires to occur, it ought to be based on an event B befell already. P(B|A) is a conditional likelihood of an match B with conditioned an event A. If an event B wants to occur, it have to be based on an event A occurred already. P(B) = Prior likelihood of an match B.Here, B is now not structured on an event A in anyway.

Naive Bayes Classifier: Naive Bayes classifier considers, all the facets (attributes) of the dataset independently contribute to classify the new data, even attributes have some dependency. It means, one attribute chance has to not impact the incidence of other attribute's likelihood in the data set. Also, every attribute in the statistics set equally contributes to predict the new records category label. As per Bayes theorem P(A|B) is known as as aposterior probability. In

Naive Bayes classifier, posterior likelihood has to be calculated for all the attributes independently. Then the best posterior likelihood attributes will be taken as most likely attribute and referred to as it as Maximum posteriori (MAP).

MAP (A) = Max (P (A|B))

MAP (A) = Max ((P (B|A) * P (A))/P (B)

Here, P (B) act as proof chance with steady value and which is helps to normalize the end result only. Due to P (B) is a constant, it can be left out and it will no longer have an effect on the

MAP (A) value. So,

MAP (A) = Max ((P(B|A)*P(A)))

Types of Naive Bayes Algorithm:

1. Gaussian Naive Bayes
2. Multinomial Naive Bayes
3. Bernoulli Naive Bayes

Gaussian Naive Bayes is useful, if all the attribute values are continuous. It performs Normal Distribution and calculates the suggest and variance for all attribute values. Multinomial Naive Bayes is beneficial when attribute values are allotted multi-nominally. Bernoulli Naive Bayes is useful, when the attribute values are binary-valued.

Advantages

The followings are some execs of the use of Naïve Bayes classifiers −

- Naïve Bayes classification is effortless to put in force and fast.
- It will converge quicker than discriminative fashions like logistic regression.
- It requires much less education data.
- It is especially scalable in nature, or they scale linearly with the variety of predictors and facts points.
- It can make probabilistic predictions and can take care of non-stop as properly as discrete data.
- Naïve Bayes classification algorithm can be used for binary as properly as multi-class classification issues both.

Disadvantages

The followings are some disadvantages of the usage of Naïve Bayes classifiers −

- One of the most essential cons of Naïve Bayes classification is its robust function independence due to the fact in actual existence it is nearly not possible to have a set of facets which are totally impartial of every other.
- Another trouble with Naïve Bayes classification is its 'zero frequency' which capacity that if a categorial variable has a class however no longer being discovered in coaching facts set, then Naïve Bayes mannequin will assign a zero likelihood to it and it will be unable to make a prediction.

5. UNSUPERVISED MACHINE LEARNING

There are instances when you don't choose to exactly predict an outcome. You simply choose to perform segmentation or clustering. For example – a financial institution would desire to have a segmentation of its clients to understand their behavior. This is an Unsupervised Machine Learning problem as we are now not predicting any outcomes here.

5.1 K-MEANS CLUSTERING ALORITHM

K-means clustering algorithm computes the centroids and iterates till we it finds most beneficial centroid. It assumes that the wide variety of clusters are already known. It is additionally known as flat clustering algorithm. The wide variety of clusters recognized from facts by using algorithm is represented by way of 'K' in K-means.

In this algorithm, the records factors are assigned to a cluster in such a manner that the sum of the squared distance between the facts factors and centroid would be minimum. It is to be understood that much less version inside the clusters will lead to extra comparable information factors inside identical cluster.

We can understand the working of K-Means clustering algorithm with the assist of following steps −

Step 1 − First, we want to specify the variety of clusters, K, want to be generated through this algorithm.

Step 2 − Next, randomly choose K records factors and assign every statistics factor to a cluster. In easy words, classify the information primarily based on the range of statistics points.

Step 3 − Now it will compute the cluster centroids.

Step 4 − Next, preserve iterating the following till we locate ideal centroid which is the undertaking of facts factors to the clusters that are now not altering any more

4.1 − First, the sum of squared distance between records factors and centroids would be computed.

4.2 − Now, we have to assign every information factor to the cluster that is nearer than different cluster (centroid).

4.3 − At final compute the centroids for the clusters by using taking the common of all statistics factors of that cluster.

K-means follows Expectation-Maximization method to resolve the problem. The Expectation-step is used for assigning the records factors to the closest cluster and the Maximization-step is used for computing the centroid of every cluster. While working with K-means algorithm we want to take care of the following matters −

While working with clustering algorithms together with K-Means, it is advocated to standardize the information due to the fact such algorithms use distance-based dimension to decide the similarity between statistics points.

Due to the iterative nature of K-Means and random initialization of centroids, K-Means may additionally stick in a neighbourhood optimal and may additionally no longer converge

Advantages

- It is computationally more efficient than hierarchical clustering when variables are huge.
- With globular cluster and small k it produces tighter clusters than hierarchical clustering. Ease in implementation and interpretation of the clustering consequences are the enchantment of this algorithm.
- Order of complexity of the algorithm is O(K*n*d) and so it is computationally environment friendly

Disadvantages

- Prediction of K cost is hard. Performance suffers when clusters are globular. Also seeing that specific initial partitions end result in exceptional last clusters it impacts performance.
- Performance degrades when there is difference in the measurement and density in the clusters in the enter data.
- Uniform impact regularly produces clusters with quite uniform size even if the enter information have specific cluster size.
- Spherical assumption (i.e. joint distribution of aspects inside each cluster is spherical) is difficult to be relaxed as the correlation between aspects destroy it and would put greater weights on correlated features.
- K fee is now not known. It is touchy to outliers. It is touchy to preliminary factors and nearby optimal, and there is no special answer for a sure K fee - so one needs to run K suggest for a K cost loads of times(20-100times) and then choose the effects with lowest J.

5.2 HIERARCHICAL CLUSTERING ALGORITHM

Hierarchical clustering is every other unsupervised studying algorithm that is used to crew collectively the unlabelled information factors having comparable characteristics. Hierarchical clustering algorithms falls into following two categories.

Agglomerative hierarchical algorithms –

The Agglomerative Hierarchical Clustering is the most frequent type of hierarchical clustering used to team objects in clusters based totally on their similarity. It's additionally regarded as AGNES (Agglomerative Nesting). It's a "bottom-up" approach: every commentary begins in its very own cluster, and pairs of clusters are merged as one strikes up the hierarchy.

Steps:

- Make every fact factor a single-point cluster → varieties N clusters
- Take the two closest records factors and make them one cluster → types N-1 clusters
- Take the two closest clusters and make them one cluster → Forms N-2 clusters.
- Repeat step-3 till you are left with solely one cluster. In agglomerative hierarchical algorithms, every statistics factor is handled as a single cluster and then successively merge or agglomerate (bottom-up approach) the pairs of clusters. The hierarchy of the clusters is represented as a dendrogram or tree structure.

There are numerous approaches to measure the distance between clusters in order to determine the policies for clustering, and they are frequently referred to as Linkage Methods. Some of the frequent linkage strategies are:

Complete-linkage: the distance between two clusters is described as the longest distance between two factors in every cluster.

Single-linkage: the distance between two clusters is described as the shortest distance between two points in every cluster. This linkage may also be used to notice excessive values in your dataset which might also be outliers as they will be merged at the end.

Average-linkage: the distance between two clusters is described as the average distance between every factor in one cluster to each and every factor in the different cluster.

Centroid-linkage: finds the centroid of cluster 1 and centroid of cluster 2, and then calculates the distance between the two earlier than merging.

The desire of linkage technique completely relies upon on you and there is no challenging and quick approach that will usually provide you desirable results. Different linkage strategies lead to one-of-a-kind clusters.

The factor of doing all this is to reveal the way hierarchical clustering works, it keeps a reminiscence of how we went via this procedure and that reminiscence is saved in Dendrogram.

Divisive hierarchical algorithms

On the different hand, in divisive hierarchical algorithms, all the records factors are handled as one large cluster and the method of clustering entails dividing (Top-down approach) the one huge cluster into a number of small clusters.

In Divisive or DIANA (DIvisive ANAlysis Clustering) is a top-down clustering approach the place we assign all of the observations to a single cluster and then partition the cluster to two least comparable clusters. Finally, we proceed recursively on every cluster till there is one cluster for every observation. So, this clustering strategy is precisely contrary to Agglomerative clustering.

There is proof that divisive algorithms produce extra correct hierarchies than agglomerative algorithms in some instances however is conceptually greater complex.

In each agglomerative and divisive hierarchical clustering, customers want to specify the preferred variety of clusters as a termination condition (when to end merging).

6. REINFORCEMENT LEARNING (RL)

Reinforcement Learning (RL) describes a kind of Machine Learning method in that the agent receives a late reward in the very next time step to gauge its previous action. RL setup consists of two components, an agent and an environment. The difference between model free and model based is the model represents the simulation of the dynamics of the environment. That's, the model learns the transition probability *T(s1|(s0, a))* from the set of current state s*0* and action A to the next state s*1*.If the transition probability is successfully learned, the agent will know how likely to enter a certain state given current state and action. Be that as it may, model-based calculations become unreasonable as the state space and activity space develops (S * S * A, for a plain arrangement). On one other hand, model-free algorithms rely on trial-and-error to update its knowledge. As a result, it generally does not require space to store most of the combination of states and actions [2].

- On Policy: the learning agent learns the value function according to the current action produced from the policy currently being used.
- Off Policy: the learning agent learns the value function according to the action produced from another policy.

Q-Learning technique is an Off-Policy technique and uses the greedy approach to master the Q-value. SARSA technique, on one other hand, is an On Policy and uses the action performed by the existing policy to learn the Q-value. This difference is visible in the difference of the update statements for each technique:

6.1 Q-LEARNING:

Q-learning, utilize dynamic programming techniques. MDP introduces with the optimum policy idea to be able to accomplish probably the most extreme rewards with time. Fundamentals of RL may be listed the following:

- The agent cooperates with the environmental surroundings and takes action Aτ in each state Sτ and waits for the response.
- The environment issues a reward (Rτ) for the accomplished actions, which can have either positive (R+) or negative value (R−).
- The agent observes the environmental surroundings for almost any changes and optimize the received rewards by updating the policies. It's worth mentioning that we adopt Q-learning since the reinforcement learning method due to its model-free nature. In addition, Q-learning will attract because when applied, it has the capacity to learn without following the existing policy.

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma \max_a Q(s_{t+1}, a) - Q(s_t, a_t)]$$

Q-Learning builds on the concept of state value iteration where a real estate agent aims to estimate the state value function V (S) to update all states S and actions A for every emphasis so as to know which an outcome in higher reward R. Within the Q table, the rows represent the states whereas the columns represent what. In every S, the agent takes a motion A, watches the reward for this action R and also the following state Sˆ, and updates the estimated Q (EST − Q) value [16].

Algorithm - Q-learning

Q-learning: Learn function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

Require:
Sates $\mathcal{X} = \{1, \ldots, n_x\}$
Actions $\mathcal{A} = \{1, \ldots, n_a\}$, $A : \mathcal{X} \Rightarrow \mathcal{A}$
Reward function $R : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$
Black-box (probabilistic) transition function $T : \mathcal{X} \times \mathcal{A} \to \mathcal{X}$
Learning rate $\alpha \in [0, 1]$, typically $\alpha = 0.1$
Discounting factor $\gamma \in [0, 1]$
procedure QLEARNING($\mathcal{X}$, A, R, T, α, γ)
  Initialize $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ arbitrarily
  while Q is not converged **do**
    Start in state $s \in \mathcal{X}$
    while s is not terminal **do**
      Calculate π according to Q and exploration strategy (e.g. $\pi(x) \leftarrow \arg\max_a Q(x, a)$)
      $a \leftarrow \pi(s)$
      $r \leftarrow R(s, a)$ ▷ Receive the reward
      $s' \leftarrow T(s, a)$ ▷ Receive the new state
      $Q(s', a) \leftarrow (1 - \alpha) \cdot Q(s, a) + \alpha \cdot (r + \gamma \cdot \max_{a'} Q(s', a'))$
      $s \leftarrow s'$
  return Q

In comparison, Q-learning has no constraint over the following action, so long as it maximizes the Q-value for the following state. Therefore, SARSA can be an on-policy algorithm. The updated equation for SARSA depends upon the existing state, current action, reward obtained, next state and next action. This observation result in the naming of the educational technique as SARSA stands for **State Action Reward State Action** which symbolizes the tuple (s, a, r, s ', a').

6.2 SARSA ALGORITHM

SARSA algorithm is really a slight variation of the most popular Q-Learning algorithm. **SARSA** (whose name is derived from the sequence *state-action-reward-state-action*) is an all-natural extension of TD (0) to the estimation of the *Q* function. Its standard formulation (which may also be called one-step SARSA, or SARSA (0), is based on an individual next reward, r_{t+1}, which will be obtained by executing the action a_t in the state s_t. The temporal difference computation is on the basis of the following update rule:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha[r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)]$$

The action a_(t+1) may be the action performed next state S(t+1) under current policy.

SARSA(λ): Learn function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$

Require:
Sates $\mathcal{X} = \{1, \ldots, n_x\}$
Actions $\mathcal{A} = \{1, \ldots, n_a\}$, $A : \mathcal{X} \Rightarrow \mathcal{A}$
Reward function $R : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$
Black-box (probabilistic) transition function $T : \mathcal{X} \times \mathcal{A} \to \mathcal{X}$
Learning rate $\alpha \in [0, 1]$, typically $\alpha = 0.1$
Discounting factor $\gamma \in [0, 1]$
$\lambda \in [0, 1]$: Trade-off between TD and MC
procedure QLEARNING($\mathcal{X}$, A, R, T, α, γ, λ)
  Initialize $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ arbitrarily
  Initialize $e : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ with 0. ▷ eligibility trace
  while Q is not converged **do**
    Select $(s, a) \in \mathcal{X} \times \mathcal{A}$ arbitrarily
    while s is not terminal **do**
      $r \leftarrow R(s, a)$
      $s' \leftarrow T(s, a)$ ▷ Receive the new state
      Calculate π based on Q (e.g. epsilon-greedy)
      $a' \leftarrow \pi(s')$
      $e(s, a) \leftarrow e(s, a) + 1$
      $\delta \leftarrow r + \gamma \cdot Q(s', a') - Q(s, a)$
      for $(\tilde{s}, \tilde{a}) \in \mathcal{X} \times \mathcal{A}$ **do**
        $Q(\tilde{s}, \tilde{a}) \leftarrow Q(\tilde{s}, \tilde{a}) + \alpha \cdot \delta \cdot e(\tilde{s}, \tilde{a})$
        $e(\tilde{s}, \tilde{a}) \leftarrow \gamma \cdot \lambda \cdot e(\tilde{s}, \tilde{a})$
      $s \leftarrow s'$
      $a \leftarrow a'$
  return Q

On the other hand, Q-learning has no imperative throughout the following activity, as long as it amplifies the Q-esteem for the following state. Therefore, SARSA is an on-policy algorithm. The updated equation for SARSA depends on the current state, current action, reward obtained, next state and next action. This observation lead to the naming of the learning technique as SARSA stands for **State Action Reward State Action** which symbolizes the tuple (s, a, r, s', a') [17].

6.3 DEEP Q NEURAL NETWORK (DQN)

DQN is Q-learning with Neural Networks. The inspiration driving is just identified with enormous state space conditions where characterizing a Q-table would be a mind boggling, testing and tedious assignment. Rather than a Q-table Neural Networks rough Q-values for each activity dependent on the state.

7. CONCLUSION

This paper presents the view of Machine Learning algorithms to Big Data context. Next, the technique of Machine Learning algorithms processing and analysing Big Data is identified and addressed. All this, allowed us to succeed the task of adapting Machine Learning algorithms to the Big Data context. The algorithm that efficiently used in big data analytics are discusses. In supervised algorithm classification, Regression, Decision Tree, KNN, Random Forest, SVM, Naive Bayes are discussed. In Unsupervised algorithm K-Means Clustering, Hierarchical Clustering algorithms are discussed. In reinforcement algorithm Q-learning, SARSA, Deep Q Neural Network (DQN) are discussed. In future these algorithms must expand for better performance in big data analytics. There are many points to be considered, discussed, improved, developed, analysed, etc. Hopefully it has provided some useful discussion and a structure for researchers.

REFERENCES

1. Dwivedi, Y. K., Janssen, M., Slade, E. L., Rana, N. P., Weerakkody, V., Millard, J., Snijders, D. (2017). Driving innovation through big open linked data (BOLD): Exploring antecedents using interpretive structural modeling. Information Systems Frontiers, 19(2), 197-212. https://doi.oig/10.1007/sl0796-016-9675-5
2. SASSI, S. OUAFTOUH and S. ANTER, "Adaptation of Classical Machine Learning Algorithms to Big Data Context: Problems and Challenges: Case Study: Hidden Markov Models Under Spark," 2019 1st International Conference on Smart Systems and Data Science (ICSSD), Rabat, Morocco, 2019, pp. 1-7, doi: 10.1109/ICSSD47982.2019.9002857.
3. 4.Tang, Yan & Wang, Yu & Cooper, K.M.L. & Li, Ling. (2014). Towards Big Data Bayesian Network Learning - An Ensemble Learning Based Approach. Proceedings - 2014 IEEE International Congress on Big Data, BigData Congress 2014. 355-357. 10.1109/BigData.Congress.2014.58.
4. 3.Y. Zhao-hong, W. Hui-yu, Z. Bin, H. Zhi-he and L. Wan-lin, "A literature review on the key technologies of processing big data," 2018 IEEE 3rd International Conference on Cloud Computing and Big Data Analysis (ICCCBDA), Chengdu, 2018, pp. 202-208, doi: 10.1109/ICCCBDA.2018.8386512.
5. M. D. A. Praveena and B. Bharathi, "A survey paper on big data analytics," 2017 International Conference on Information Communication and Embedded Systems (ICICES), Chennai, 2017, pp. 1-9, doi: 10.1109/ICICES.2017.8070723.
6. H. Y. Putra, H. Putra and N. B. Kurniawan, "Big Data Analytics Algorithm, Data Type and Tools in Smart City: A Systematic Literature Review," 2018 International Conference on Information Technology Systems and Innovation (ICITSI), Bandung - Padang, Indonesia, 2018, pp. 474-478, doi: 10.1109/ICITSI.2018.8696051.
7. S. Bahri, N. Zoghlami, M. Abed and J. M. R. S. Tavares, "BIG DATA for Healthcare: A Survey," in IEEE Access, vol. 7, pp. 7397-7408, 2019, doi: 10.1109/ACCESS.2018.2889180.
8. Harshawardhan S. Bhosale, Devender P. Gadekar," A Review paper on Big Data and Hadoop", International Journal of Scientific and Research Publication, Vol 4, 2014
9. Cheikh Kacfah Emani, Nadine Cullot, Christopher, Nicolle, "Understandable Big Data: A Survey", Elsevier Computer Science Review (2015)
10. S. Vikram Phannendra,E. Madhusudhana Reddy,"Big Data – Solution for RDBMS Problems- A Survey", IJARCCE(International Journal Advance Research in computer and Communication Engineering), Vol 2 (2013).
11. Vibha Shukla, Pawan Kumar Dubey," Big Data: Moving Forward with Emerging Technology and Challenges", International Journal Advance Research in computer Science and Management Studies, Vol 2(2014)
12. Prity Vijay and Brigh Keshwani," Emergence of Big Data and Hadoop: A Review", IOSR Journal of Engineerinf, Vol 6(2016)
13. Raymond Gardiner Goss," Heading Towards Big Data ", IEEE (2013)
14. Poonam S. Patil and Rajesh N. Phursule," Survey Paper on Big Data and Hadoop Component", International Journal of Scientific and Research, Vol 3, 2014
15. Kuchipudi Sravanthi, T. Subba Reddy," Application of Big Data in Various Fields", International Journal of computer Science and Information Technologies, Vol 6(2015)
16. 1.S. Otoum, B. Kantarci and H. Mouftah, "Empowering Reinforcement Learning on Big Sensed Data for Intrusion Detection," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 2019, pp. 1-7, doi: 10.1109/ICC.2019.8761575.
17. Kung-Hsiang, Huang (Steeve), Co-Founder & CTO, Rosetta.ai | Research Assistant, ISI |
18. J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques", 3rd Edition, MK Series, 2012.

19. Susmita Ray," A Quick Review of Machine Learning Algorithms", 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (Com-IT-Con), India, 14th -16th Feb 2019
20. S. Otoum, B. Kantarci and H. Mouftah, "Empowering Reinforcement Learning on Big Sensed Data for Intrusion Detection," ICC 2019 - 2019 IEEE International Conference on Communications (ICC), Shanghai, China, 2019, pp. 1-7, doi: 10.1109/ICC.2019.8761575.
21. Dr. O. Obulesu, M. Mahendra, M. ThrilokReddy "Machine Learning Techniques and Tools: A Survey", 2018 Proceedings of the International Conference on Inventive Research in Computing Applications (ICIRCA 2018)
IEEE Xplore Compliant Part Number:CFP18N67-ART;ISBN:978-1-5386-2456-2