IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

Big Data Analytics Algorithm, Tools In Systematic Review

M. THARANIDEVI,
Asst.Prof.,
Nadar Sarawathi College of Arts and Science, Theni

Abstract : ***Big data analytics describes the process of uncovering trends, patterns, and correlations in large amounts of raw data to help make data-informed decisions. These processes use familiar statistical analysis techniques—like clustering and regression—and apply them to more extensive datasets with the help of never tools. new technologies—from Amazon to smartphones—have contributed even more to the substantial amounts of data available to organizations. Five fundamentals algorithm can review on this paper. There are hundreds of data analytics tools out there in the market today but the selection of the right tool will depend upon your business NEED, GOALS, and VARIETY to get business in the right direction.***

Keynotes: Introduction of Big Data Analytics, Fundamentals of Algorithms and Tools

INTRODUCTION

Big data analytics is the often complex process of examining big data to uncover information. Such as hidden patterns, correlations, market trends and customer preferences—that can help organizations make informed business decisions. This data analytics is a form of advance analytics, which involve complex applications with elements such as predictive models, statistical algorithms and what -if analysis powered by analytics systems. This data analytics can used to most powerful algorithm.

FUNDAMENTALS OF ALGORITHMS

Algorithm: An Algorithm is a procedure used for solving a problem or performing a computation. This paper can explain 4 types of algorithm.

1. Linear Regression
2. Logistic Regression
3. K-Nearest Neighbour
4. K-Means Clustering

1.Linear Regression:

Linear regression is one of the most basic algorithms of advanced analytics. This also makes it one of the most widely used. People can easily visualize how it is working and how the input data is related to the output data.

Linear regression uses the relationship between two sets of continuous quantitative measures. The first set is called the *predictor* or *independent variable*. The other is the *response* or *dependent variable*. The goal of linear regression is to identify the relationship in the form of a formula that describes the dependent variable in terms of the independent variable. Once this relationship is quantified, the dependent variable can be predicted for any instance of an independent variable.

One of the most common independent variables used is time. Whether your independent variable is revenue, costs, customers, use, or productivity, if you can define the relationship it has with time, you can forecast a value with linear regression.

2. Logistic Regression:

Logistic regression sounds similar to linear regression but is actually focused on problems involving categorization instead of quantitative forecasting. Here the output variable values are discrete and finite rather than continuous and with infinite values as with linear regression.

The goal of logistic regression is to categorize whether an instance of an input variable either fits within a category or not. The output of logistic regression is a value between 0 and 1. Results closer to 1 indicate that the input variable more clearly fits within the category. Results closer to 0 indicate that the input variable likely does not fit within the category.

Logistic regression is often used to answer clearly defined yes or no questions. Will a customer buy again? Is a buyer credit worthy? Will the prospect become a customer? Predicting the answer to these questions can spawn a series of actions within the business process which can help drive future revenue.

Classification and Regression Trees

Classification and regression trees use a decision to categorize data. Each decision is based on a question related to one of the input variables. With each question and corresponding response, the instance of data gets moved closer to being categorized in a specific way. This set of questions and responses and subsequent divisions of data create a tree-like structure. At the end of each line of questions is a category. This is called the *leaf node* of the classification tree.

These classification trees can become quite large and complex. One method of controlling the complexity is through pruning the tree or intentionally removing levels of questioning to balance between exact fit and abstraction. A model that works well with all instances of input values, both those that are known in training and those that are not, is paramount. Preventing overfitting of this model requires a delicate balance between exact fit and abstraction.

A variant of classification and regression trees is called *random forests*. Instead of constructing a single tree with many branches of logic, a random forest is a culmination of many small and simple trees that each evaluate the instances of data and determine a categorization. Once all of these simple trees complete their data evaluation, the process merges the individual results to create a final prediction of the category based on the composite of the smaller categorizations. This is commonly referred to as an *ensemble method*. These random forests often do well at balancing exact fit and abstraction and have been implemented successfully in many business cases.

In contrast to logistic regression, which focuses on a yes or no categorization, classification and regression trees can be used to predict multivalue categorizations. They are also easier to visualize and see the definitive path that guided the algorithm to a specific categorization.

3. K-Nearest Neighbour:

K-nearest neighbor is also a classification algorithm. It is known as a "lazy learner" because the training phase of the process is very limited. The learning process is composed of the training set of data being stored. As new instances are evaluated, the distance to each data point in the training set is evaluated and there is a consensus decision as to which category the new instance of data falls into based on its proximity to the training instances.

This algorithm can be computationally expensive depending on the size and scope of the training set. As each new instance has to be compared to all instances of the training data set and a distance derived, this process can use many computing resources each time it runs.

This categorization algorithm allows for multivalued categorizations of the data. In addition, noisy training data tends to skew classifications.

K-nearest neighbors is often chosen because it is easy to use, easy to train, and easy to interpret the results. It is often used in search applications when you are trying to find similar items.

4.K-Means Clustering:

K-means clustering focuses on creating groups of related attributes. These groups are referred to as clusters. Once these clusters are created, other instances can be evaluated against them to see where they best fit.

This technique is often used as part of data exploration. To start, the analyst specifies the number of clusters. The K-means cluster process breaks the data into that number of clusters based on finding data points with similarities around a common hub, called the *centroid*. These clusters are not the same as categories because initially they do not have business meaning. They are just closely related instances of input variables. Once these clusters are identified and analyzed, they can be converted to categories and provided a name that has business meaning.

K-means clustering is often used because it is simple to use and explain and because it is fast. One area to note is that k-means clustering is extremely sensitive to outliers. These outliers can significantly shift the nature and definition of these clusters and ultimately the results of analysis.

These are some of the most popular algorithms in use in advanced analytics initiatives. Each has pros and cons and different ways in which it can be effectively utilized to generate business value. The end target with the implementation of these algorithms is to further refine the data to a point where the information that results can be applied to business decisions. It is this process of informing downstream processes with more refined and higher value data that is a fundamental to companies becoming truly harnessing the value of their data and achieving the results that they desire.

Big data analytic Tools:

More than Tools available there still this paper can explain some of the tools.

1. APACHE Hadoop
2. Cassandra
3. Qubole
4. Xplenty
5. Spark
6. Mongo DB
7. Apache Storm
8. SAS
9. Data Pine
10. Rapid Miner

These are explain to below.

1.APACHE Hadoop

It's a Java-based open-source platform that is being used to store and process big data. It is built on a cluster system that allows the system to process data efficiently and let the data run parallel. It can process both structured and unstructured data from one server to multiple computers. Hadoop also offers **cross-platform** support for its users. Today, it is the best ***big data analytic tool*** and is popularly used by many tech giants such as Amazon, Microsoft, IBM, etc.

Features of Apache Hadoop:

- Free to use and offers an efficient storage solution for businesses.
- Offers quick access via HDFS (Hadoop Distributed File System).
- Highly flexible and can be easily implemented with MySQL, and JSON.
- Highly scalable as it can distribute a large amount of data in small segments.
- It works on small commodity hardware like JBOD or a bunch of disks.

2.Cassandra

APACHE Cassandra is an open-source NoSQL distributed database that is used to fetch large amounts of data. It's one of the **most popular tools for data analytics** and has been praised by many tech companies due to its high scalability and availability without compromising speed and performance. It is **capable of delivering thousands of operations every second** and can handle petabytes of resources with almost zero downtime. It was created by Facebook back in 2008 and was published publicly.

Features of APACHE Cassandra:

- *Data Storage Flexibility:* It supports all forms of data i.e. structured, unstructured, semi-structured, and allows users to change as per their needs.
- *Data Distribution System:* Easy to distribute data with the help of replicating data on multiple data centers.
- *Fast Processing:* Cassandra has been designed to run on efficient commodity hardware and also offers fast storage and data processing.

Fault-tolerance: The moment, if any node fails, it will be replaced without any delay.

3.Qubole

It's an open-source big data tool that helps in fetching data in a value of chain using ad-hoc analysis in machine learning. Qubole is a data lake platform that offers end-to-end service with reduced time and effort which are required in moving data pipelines. It is capable of configuring multi-cloud services such as AWS, Azure, and Google Cloud. Besides, it also helps in lowering the cost of cloud computing by 50%.

Features of Qubole:

- *Supports ETL process:* It allows companies to **migrate data from multiple sources in one place**.
- *Real-time Insight:* It monitors user's systems and allows them to view real-time insights
- *Predictive Analysis:* Qubole offers predictive analysis so that companies can take actions accordingly for targeting more acquisitions.
- *Advanced Security System:* To protect users' data in the cloud, Qubole uses an advanced security system and also ensures to protect any future breaches. Besides, it also allows encrypting cloud data from any potential threat.

4.Xplenty

It is a data analytic tool for building a data pipeline by using minimal codes in it. It offers a wide range of solutions for sales, marketing, and support. With the help of its interactive graphical interface, it provides solutions for *ETL*, *ELT*, etc. The best part of using Xplenty is its low investment in hardware & software and its offers support via **email, chat, telephonic and virtual meetings**. Xplenty is a platform to process data for analytics over the cloud and segregates all the data together.

Features of Xplenty:

- *Rest API:* A user can possibly do anything by implementing Rest API
- *Flexibility:* Data can be sent, and pulled to databases, warehouses, and salesforce.
- *Data Security:* It offers SSL/TSL encryption and the platform is capable of verifying algorithms and certificates regularly.
- *Deployment:* It offers integration apps for both cloud & in-house and supports deployment to integrate apps over the cloud.

***5.*Spark**

APACHE Spark is another framework that is used to process data and perform numerous tasks on a large scale. It is also used to process data via multiple computers with the help of distributing tools. It is widely used among data analysts as it offers easy-to-use APIs that provide easy data pulling methods and it is **capable of handling multi-petabytes of data** as well. Recently, Spark made a record of processing **100 terabytes of data in just 23 minutes** which broke the previous world record of **Hadoop (71 minutes)**. This is the reason why big tech giants are moving towards spark now and is highly suitable for ML and AI today.

Features of APACHE Spark:

- *Ease of use:* It allows users to run in their preferred language. (JAVA, Python, etc.)
- *Real-time Processing:* Spark can handle real-time streaming via Spark Streaming
- *Flexible:* It can run on, Mesos, Kubernetes, or the cloud.

6.Mongo DB

Came in limelight in 2010, is a free, open-source platform and a **document-oriented (NoSQL) database** that is used to store a high volume of data. It uses collections and documents for storage and its document consists of key-value pairs which are considered a basic unit of Mongo DB. It is so popular among developers due to its availability for multi-programming languages such as Python, Jscript, and Ruby.

Features of Mongo DB:

- *Written in C++:* It's a schema-less DB and can hold varieties of documents inside.
- *Simplifies Stack:* With the help of mongo, a user can easily store files without any disturbance in the stack.
- *Master-Slave Replication*: It can write/read data from the master and can be called back for backup.

***7.*Apache Storm**

A storm is a robust, user-friendly tool used for data analytics, especially in small companies. The best part about the storm is that it has no language barrier (programming) in it and can support any of them. It was designed to **handle a pool of large data in fault-tolerance and horizontally scalable methods**. When we talk about real-time data processing, Storm leads the chart because of its distributed real-time big data processing system, due to which today many tech giants are using APACHE Storm in their system. Some of the most notable names are Twitter, Zendesk, NaviSite, etc.

Features of Storm:

- *Data Processing:* Storm process the data even if the node gets disconnected
- *Highly Scalable:* It keeps the momentum of performance even if the load increases
- *Fast:* The speed of APACHE Storm is impeccable and can process up to 1 million messages of 100 bytes on a single node.

***8.*SAS**

Today it is one of the best tools for creating statistical modeling used by data analysts. By using SAS, a data scientist can mine, manage, extract or update data in different variants from different sources. Statistical Analytical System or SAS allows a user to access the data in any format (SAS tables or Excel worksheets). Besides that it also offers a cloud platform for business analytics called **SAS Viya** and also to get a strong grip on AI & ML, they have introduced new tools and products.

Features of SAS:

- *Flexible Programming Language:* It offers easy-to-learn syntax and has also vast libraries which make it suitable for non-programmers
- *Vast Data Format:* It provides support for many programming languages which also include SQL and carries the ability to read data from any format.
- *Encryption:* It provides end-to-end security with a feature called **SAS/SECURE**.

***9.*Data Pine**

Datapine is an analytical used for BI and was founded back in 2012 (Berlin, Germany). In a short period of time, it has gained much popularity in a number of countries and it's mainly used for data extraction (for small-medium companies fetching data for close monitoring). With the help of its enhanced UI design, anyone can visit and check the data as per their requirement and offer in 4 different price brackets, starting from $249 per month. They do offer dashboards by functions, industry, and platform.

Features of Datapine:

- *Automation:* To cut down the manual chase, datapine offers a wide array of AI assistant and BI tools.
- *Predictive Tool:* datapine provides forecasting/predictive analytics by using historical and current data, it derives the future outcome.
- *Add on:* It also offers intuitive **widgets, visual analytics & discovery, ad hoc reporting**, etc.

***10.*Rapid Miner**

It's a fully automated visual workflow design tool used for data analytics. It's a no-code platform and users aren't required to code for segregating data. Today, it is being heavily used in many industries such as ed-tech, training, research, etc. Though it's an open-source platform but has a limitation of adding **10000 data rows and a single logical processor**. With the help of Rapid Miner, one can easily deploy their ML models to the web or mobile (only when the user interface is ready to collect real-time figures).

Features of Rapid Miner:

- *Accessibility:* It allows users to access 40+ types of files (SAS, ARFF, etc.) via URL
- *Storage:* Users can access cloud storage facilities such as AWS and dropbox
- *Data validation:* Rapid miner enables the visual display of multiple results in history for better evaluation.

Conclusion:

Big Data is a data process that exceeds the capacity to be processed in conventional databases, so it is necessary to use technology for extensive data processing and fast processing in helping companies deal with data errors, from various analyzes that have described that big data has many benefits to explore the potential for a set of data to be reused in addressing and increasing potential problems to be resolved, big data has made it easier to make decisions.

Reference:

[1] B. Daniel, "Big Data and analytics in higher education: Opportunities and challenges," Br. J. Educ. Technol., 2015.

[2] S. Fosso Wamba, S. Akter, A. Edwards, G. Chopin, and D. Gnanzou, "How 'big data' can make big impact: Findings from a systematic review and a longitudinal case study," Int. J. Prod. Econ., 2015.

[3] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," J. Bus. Res., 2017.

[4] E. Al Nuaimi, H. Al Neyadi, N. Mohamed, and J. Al-Jaroodi, "Applications of big data to smart cities," J. Internet Serv. Appl., 2015.

[5] D. Singh and C. K. Reddy, "A survey on platforms for big data analytics," J. Big Data, 2015. [6] M. Cox and D. Ellsworth, "Managing Big Data for Scientific Visualization," ACM Siggraph,1997