International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 11 | Nov 2019 www.irjet.net p-ISSN: 2395-0072

A Survey on Predictive Analytics and Parallel Algorithms for Knowledge Extraction from Data Received through Various Satellites

Bharani B R[1], Dr Gururaj Murtugudde[2]

[1]*Assisstant Professor, Cambridge Institute of Technology, Bangalore, India.*
[2]*Prof & Head, Dept of CSE, Nagarjuna College of Engineering and Technology, Bangalore, India.*

Abstract:- *The remote sensing satellites produce large volumes of data that cannot be stored in standard relational databases every day. Many software components extract information in unstructured form from the raw data producing information such as pictures, log files, pdf user instructions, word etc. there is a need for developing efficient data mining algorithms to tag the datasets for facilitating efficient buildup of archival and retrieval. Advances in remote sensing instruments and technology are transforming the way satellite data is collected, managed and analyzed[1]. Recently, efforts have been directed towards knowledge extraction and analysis of satellite data[9]. However, the approach poses complex computational problem in terms of processing huge volume of varied form of data[8]. Still, many current and future satellite applications require the incorporation of Apache Spark and Hadoop Distributed File Systems(HDFS) technologies with real time processing capabilities. SQL database servers have traditionally held gigabytes of information. In the past 15 years, data warehouses and enterprise analytics expanded these volumes to terabytes. In the last five years, the distributed file systems that store big data now routinely house petabytes of information. This paper presents a comparative study of the data storage techniques and the different Apache tools used for data storage and the methodologies to incorporate them[7].*

Key Words: **Apache Spark, Apache Hadoop, Big Data, Remote Sensing, Knowledge Extraction**

1. INTRODUCTION

Big data is applied to information sets whose size or type goes beyond traditional relational databases' capacity to collect, handle and process low latency information. It has one or more of the following features - high volume, high speed, high range. Big data is available from sensors, computers, video/audio, networks, log files, transactional applications, internet and social media, much of it is produced in actual time and on a very big scale. These information is collected in extraordinarily growing databases that are complex to contain, form, store, handle, share, process, evaluate and visualize through typical software instruments for databases. Continuous high velocity data stream or offline high volume data to "Big Data" brings us to a new challenge[2]. Big data enables analysts, scientists and company users to make better and quiker choices using previously inaccessible or unusable information. Using sophisticated analytical methods such as text analysis, machine learning, predictive analytics, information mining, statistics and natural linguistic processing, companies can evaluate earlier untapped information sources independently or in conjunction with current business information to obtain new ideas resulting in considerably better and quicker choices. Predictive analytics is a collective word for methods intended to predict the future on the basis of static or historical information. Techniques will be used in the areas of statistics and machine learning. A predictive analysis engine or program of forecasting will comprise models of regression and/or machine learning neural networks. In predictive analysis, the concept of a model is crucial; the model determines the data based prediction. This model is constantly adapted, tuned, optimized and educated in accordance with the setting and changing user perspectives. The most effective prediction algorithms are kmeans, decision trees, rule based classifiers, deep learning and random forests. The data is aquired from satellites and collected in the base stations in the earth[1][3][4][5]. The received data is processed using the HDFS and Apache Spark tools. The processing of data involves cleaning the data and transforming the data in the required format. Then knowledge, patterns, trends are extracted from the processed data using machine learning algorithms[6][10]. The patterns and trends extracted are presented in the graphical form. This paper lists the different methodologies to process the data, different data sources for sourcing the data and machine learning algorithms.

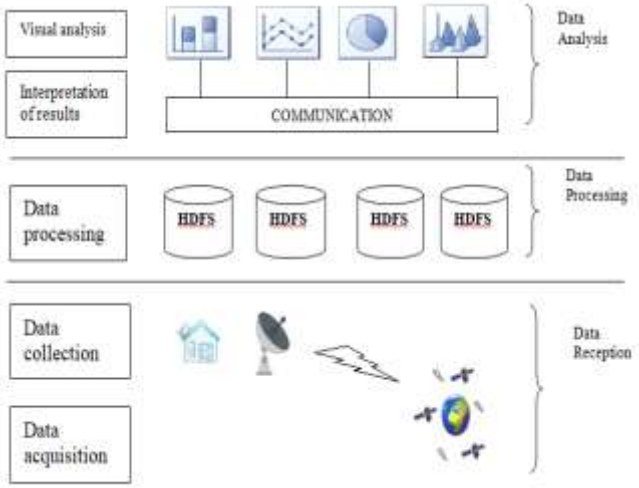

Fig -1: Satellite Data Processing

Data Sets

The data is sourced from European Space Agency(ESA), Kaggle, ISRO Data Archive, NASA Data Archive, National Oceanic and Atmospheric Administration(NOAA), Indian Space Science Data Center(ISSDC), Institute of Remote Sensing and Digital Earth, Satellite Central Data Repository(SCDR).

Methodology

Apache Spark

Apache spark is an open-source distributed general purpose cluster computing framework. Spark offers an interface with implicit data parallelism and fault tolerance for programming entire clusters. Apache Spark has the Resilient Distributed Data set(RDD) as its architectural base, a readonly multiset of data objects spread across a cluster of computers, which is managed in a fault tolerant manner. Spark enables the implementation of both iterative algorithms accessing their data set multiple times in a loop, as well as interactive/ exploratory data analysis, i.e. repetitive data base style querying.

Apache Hadoop

Apache Hadoop is a collection of open source software tools that make it easy to use various computer networks to solve problems involving massive amounts of data and computation. Using the MapReduce programming model, it offers a software framework for distributed storage and processing of large data. Apache Hadoop's core consists of a storage component, known as the Hadoop Distributed File System(HDFS) and a processing part which is MapReduce programming model. Hadoop partitions files into large blocks and distributes them into clusters across nodes.

Machine Learning Algorithms

Linear regression

This is one of the approaches of the regression analysis[11][12][13]. It is used to model the relationship between one dependent variable and one or more independent variables. The data is modeled to fit a straight line. For a set of input variables(x) that are used to determine an output variable(y). A relationship exists between the input variable and output variable. In linear regression, the relationship between the input variable(x) and the output variable(y) is expressed as an equation of the form *y=ax+b*. The coefficients a and b are called regression coefficients. The regression coefficients a and b represent the slope of the line and y represents the intercept.

Logistic regression

Logistic regression[14][15][16] is used when the output belongs to a certain class or event for example pass/fail, yes/no, 0/1, healthy/not healthy, win/lose. The linear regression function is used by the logistic regression to calculate the value of dependent variable. The linear regression function is used by the logistic regression to calculate the value of dependent variable. The linear regression function is as follows.

Pred(X) = a+b1x1+b2x2+.....+ bnxn

The output obtained by the above equation is a real number, hence the usage of sigmoid to transform the output into probability value as shown below.

$$\text{Prob}(X) = \frac{1}{1 + e^{-(a+b_1x_1+b_2x_2+.....+\ b_nx_n)}}$$

The value obtained from the above equation for Prob(x) is in between 0 and 1, this value can be considered as the probability for an outcome.

CART

Classification and regression trees[17][18][19]. This algorithm follows a top-down approach. The basic algorithm used is Hunt's algorithm. It was developed by Bremman in 1984. The types of trees are Classification and regression trees. The serial implementation is tree-growth and tree-pruning. The type of data used is discrete and continuous. The types of splits are binary splits and clever surrogate splits to reduce tree depth. The splitting criteria used is Gini's coefficient. The pruning criteria is to remove the weakest links first.

KNN

K Nearest Neighbours[20][21]. This algorithm is used for classification and regression. The training data is priorly known. Given a set of test data, the data is classified based on its nearest distance to the training data. The distance is measured by Euclidean Distance.

Unsupervised Learning Algorithms

Apriori

This is the most common association rule mining algorithm used. The goal is to discover subsets prevalent to at least a minimum amount of itemsets[22][23]. A common set of items is a set of items whose support is greater than or equal to the threshold of minimum support. The property of Apriori is a downward closure property, meaning that any subsets of a frequent itemset are also frequent itemsets. Thus, if(school, college, university, campus) is a frequent itemset, then any subset such as(school, college, university) or (college, campus) are also frequent itemsets. It utilizes a bottom-up strategy; and there is a gradual increase in the size of frequent subsets, from one itemset to two itemsets, then

three itemset subsets, etc. Candidate groups at each stage are screened for minimum support against the information.

K-means

K-means is the most common algorithm for clustering[24][25]. It calculates the clusters and their centroids iteratively. Top-down strategy is used for clustering. Beginning with a number of clusters of K. this will create random centroids as starting points for cluster centers.

PCA

Principal Component Analysis[26][27] converts the original variables into a lesser set of linear combinations. The primary concept of the PCA is to decrease the dimensionality of information set composed of many correlated factors. The transformed new set of variables are called as the principal components. The principal components are the Eigen vectors of a covariance matrix. It is a technique of summarizing data. A principal component can be described as a linear combination of observed factors that are optimally weighted.

Table -1: Advantages & Disadvantages of Machine Learning Algorithms

Advantages & Disadvantages of Algorithms			
Sl No	Algorithm	Advantages	Disadvantages
1	Linear Regression	**a.** This is a straight forward algorithm capable of mapping n-dimensional data to 1-dimensional data. **b.** It operates well if there is definite linear trend in your information	**a.** The primary restriction of this algorithm is the need for linear mapping. **b.** Only n-dimensional data can be mapped to 1-dimensional data.
2	Logistic Regression	**a.** Effective method to train and predict **b.** Effective for small datasets. **c.** Simple to understand.	**a.** Not very precise **b.** Not applicable for non-linear data and complex dataset **c.** Ends up in overfitting
3	CART	**a.** It is non parametric **b.** The variable selection is performed automatically. **c.** The missing values are handled well. **d.** It is not sensitive to outliers.	a. Hard to interpret for large dataset. b. Instability of the model.
4	K Nearest Neighbour (KNN)	**a.** Robust to noisy training data. **b.** Efficient for large datasets.	**a.** The value of k should be known. **b.** It is confusing to choose the distance measure. **c.** The cost involved for the computation is high.
5	Naive Bayes	**a.** It can produce better outcomes than other classifiers for issues with a small quantity of training data because it has a low tendency to overfit **b.** It is fast to predict a new data point. **c.** CPU usage is very limited.	**a.** Performance is sensitive to misleading information. **b.** Feature interactions cannot be incorporated.
6	Apriori	**a.** It helps to reduce the candidate itemsets size. **b.** It encourages the pruning process.	**a.** The performance time is more. **b.** The computational cost is too high.
7	K-Means	**a.** Simple to implement. **b.** Scales to large datasets. **c.** Generalizes clusters, such as elliptical clusters of distinct shapes and sizes.	**a.** Choosing K manually. **b.** Dependent on initial values. **c.** Clustering data of varying sizes and density and outliers.

3. CONCLUSION

Large satellite applications are known to be standard data-intensive issues overwhelmed by huge data. Big data problems include the complexity of storing massive complex satellite data, data access patterns which are irregular, multilevel storage hierarchy management of the data, scheduling of large amounts of dependent tasks. There is no question that the current technologies and frameworks are so minimal that can fully solve the big data issues. As we have discussed above, the techniques and the methodologies of the same can solve the above issues and provide the archival of the data and retrieval of required patterns and trends which are useful for the organisation.

REFERENCES

[1] Muhammad Mazhar Ullah Rathore, Anand Paul, Awais AhmadBo-Wei Chen, Bormin Huang, and Wen Ji. "Real-Time Big Data Analytical Architecture for Remote Sensing Application". IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2015.

[2] Sreenivas R. Sukumar. "Open Research Challenges with Big Data - A Data-Scientist's Perspective". IEEE International Conference on Big Data (Big Data), 2015.

[3] Mingmin Chi, Antonio Plaza, Jo´n Atli Benediktsson, Zhongyi Sun, Jinsheng Shen, and Yangyong Zhu. " Big Data for Remote Sensing: Challenges and Opportunities, IEEE, 2016.

[4] Jun Shao, Daqi Xu, Chun Feng and Mingmin Chi. " Big Data Challenges in China Centre for Resources Satellite Data and Application"

[5] Levente J Klein, Fernando J Marianno, Conrad M Albrecht, Marcus Freitag, Siyuan Lu, Nigel Hinds, Xiaoyan Shao, Sergio Bermudez Rodriguez1 , Hendrik F Hamann. "PAIRS: A scalable geo-spatial data analytics platform". IEEE International Conference on Big Data (Big Data), 2015.

[6] Ivens Portugal, Paulo Alencar, Donald Cowan. "A Preliminary Survey on Domain-Specific Languages for Machine Learning in Big Data". 2016 IEEE International Conference on Software Science, Technology and Engineering, 2016.

[7] A.ELomari, A.MAIZATE, L.Hassouni. "Data storage in Big Data Context: A Survey". IEEE, 2016.

[8] Ujwala M. Bhangale , Kuldeep R. Kurte , Surya S. Durbha , Roger L. King , Nicolas H. Younan. "Big Data Processing Using Hpc For Remote Sensing Disaster Data". IEEE, 2016.

[9] Leland Pierce. "Big Data Issues For Remote Sensing: Variety". IEEE, 2016.

[10] Wint Nyein Chan, Thandar Thein. "A Comparative Study of Machine Learning Techniques for Real-time Multi-tier Sentiment Analysis". 1st IEEE International Conference on Knowledge Innovation and Invention 2018.

[11] Saritha K, Sajimon Abraham. "Prediction with Partitioning: Big Data Analytics Using Regression Techniques".2017 International Conference on Networks & Advances in Computational Technologies (NetACT).

[12] Ahmed Yousuf Saber, AKM Rezaul Alam. "Short Term Load Forecasting using Multiple Linear Regression for Big Data". 2017 IEEE.

[13] *Fu Zu-feng.*"Linear Regression Protocol for Privacy Protect". 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics.

[14] B. Pavlyshenko. "Machine Learning, Linear and Bayesian Models for Logistic Regression in Failure Detection Problems". 2016 IEEE International Conference on Big Data.

[15] Preeti K. Dalvi, Siddhi K. Khandge, Ashish Deomore, Aditya Bankar, Prof. V. A. Kanade. "Analysis of Customer Churn Prediction in Telecom Industry using Decision Trees and Logistic Regression". 2016 Symposium on Colossal Data Analysis and Networking (CDAN).

[16] Anjuman Prabhat, Vikas Khullar . "Sentiment classification on Big Data using Naïve Bayes and Logistic Regression". 2017 International Conference on Computer Communication and Informatics (*ICCCI* -2017).

[17] Joseph Finkelstein, In cheol Jeong. "Using CART for Advanced Prediction of Asthma Attacks Based on Telemonitoring Data". 2016 IEEE.

[18] Dr.Neeraj Bhargava, Sonia Dayma, Abishek Kumar, Pramod Singh. "An Approach for Classification using Simple CART Algorithm in Weka". 2017 11 th International Conference on Intelligent Systems and Control (ISCO).

[19] Madhav S. Vyas, Reshma Gulwani. "*Predicting Student's Performance using CART approach in Data Science".* International Conference on Electronics, Communication and Aerospace Technology ICECA 2017.

[20] Zhenjie Chen, Jingqi Yan. "Fast KNN Search for Big Data with Set Compression Tree and Best Bin First". 2016 2nd International Conference on Cloud Computing and Internet of Things (CCIOT).

[21] JI JIAQI, Yeongjee Chung. "Research on K Nearest Neighbor Join for Big Data".Proceedings of the 2017 IEEE International Conference on Information and Automation (ICIA)Macau SAR, China, July 2017.

[22] Liu Tianbiao, Hohmann Andreas. "Apriori-based Diagnostical Analysis of Passings in the Football Game". 2017 IEEE.

[23] Mohammad-Hossein Nadimi-Shahraki, Mehdi Mansouri."Hp-Apriori: Horizontal Parallel-Apriori Algorithm For Frequent Itemset Mining from Big Data". 2017 IEEE 2nd International Conference on Big Data Analysis.

[24] Fatos Xhafa, Adriana Bogza, Santi Caball, Leonard Barolli. "Apache Mahout's *k*-Means *vs*. Fuzzy *k*-Means Performance Evaluation". 2016 International Conference on Intelligent Networking and Collaborative Systems.

[25] Ilham Kusuma, M. Anwar Ma'sum, Novian Habibie, Wisnu Jatmiko, DQG Heru Suhartanto. "Design of Intelligent K-Means Based on Spark for Big Data Clustering". IWBIS 2016.

[26] T.T. Teoh, Yue Zhang,Y.Y. Nguwi, Yuval Elovici, W.L. Ng. "Analyst Intuition Inspired High Velocity Big Data Analysis Using PCA Ranked Fuzzy K-Means Clustering with Multi-Layer Perceptron (MLP) to Obviate Cyber Security Risk". 2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2017).

[27] Farhad Pourkamali-Anaraki and Stephen Becker. "Preconditioned Data Sparsification for Big Data with Applications to PCA and K-means". IEEE, 2017.

AUTHORS

Prof Bharani B R is working as Assistant Professor in Department of Information Science & Engg in Cambridge Institute of Technology. Her areas of interest are Big Data, Data Mining.

Dr Gururaj Murtugudde is the professor and Head of the Department of Computer Science & Engg, Nagarjuna College of Engg & Technology. His Areas of specialisation are Data Mining and Artificial Intelligence.