Webology (ISSN: 1735-188X)
Volume 18, Number 6, 2021

A Case Study Of Clustering And Classification Methods For Big Data Analysis In Distributed Environment

[1]Yukti Kashyap , [2]Dr. Mayank Singh Parihar

[1]Research Scholar, [2]Asst. Professor

[1,2]Dr. C. V. Raman University, Bilaspur (C.G),

Abstract: The modern age is the age of internet, communication technology and globalization which is all about innovation and highly advanced world alongside the utilizationof all the sources which are connected to the distributed environment. In the use of this technology huge amount of information is produced by many sources, many clients and many organization whom are connected through the internet in distributed environment. To store such massive volume of data many servers are used and many users can access these data from multiple servers through the connectivity. The analysis process of this huge amount of data takes lot of efforts at every level of data extraction. The key objective of this case study is to examine the analysis process done in past through the use of many different clustering and classification methods. this research article is to describe the challenges, issues and problems in big data analysis along with the tools and techniques which are used for the clustering and classification techniques. This research paper also helps to find a better way to handle the problems of big data analysis in aspects of distributed environment.

keywords: Big data analysis, Clustering and Classification, Distributed Environment.

Introduction

Distributed environment of computing provide us the facility to share information, data and resources through connectivity. In distributed computing environment information, files and data are shared among multiple computers through online access facility and they all are run as one system. In Distributed environment data are in massive volume, high in velocity and variety of information and the biggest problem of distributed environment is selection process of meaningful data from a huge data set, this huge data set or extremely large data set alsoknown as Big Data. This huge amount of data and information becomes a big challenge to verify, to access and to compute because it is very complex and time consuming. Analysis of such numerous numbers of huge data set and find out the valuable data is very complicated, costly and time consuming. To overcome from these issues we need to think that what type of data mining we wish implement, and what restrictions are placed on sharing of information, meanwhile some problems are quite tractable, others are more difficult. Data mining refers to the processof extracting useful from the database and analysis of data. Data mining primarily used to discover and indicate relationships among the data sets. Data mining technology has emerged as ameans for identifying patterns and trends from large quantities of data [1]. Analysis of numerous data in distributed environment become possible through clustering and classification methods and tools. On other hand researchers are involved to create structures,methodologies and new approaches along with suitable tool for managing, controlling and processing this volume of data, which has led to the use of data mining tools. There are two very important methods of data mining for analyzing collected data and that are classification and clustering (cluster analysis), in which classification is supervised and clustering is unsupervised method for analyzing huge data set. Classification and Clustering both have almost same working pattern and have many similarities but in classification predefined classes are used in which objects are assigned meanwhile in clustering we have to find the common things between groups of objects having similarities according to their characteristics [4]. Classification is the process of classifying the input instance based on their corresponding class labels whereas Clustering is the process of grouping the instance based on their similarities without any help of class labels. The aim of this case study is to analysis of clustering and classification tools andtechniques of data mining this will help to predict and estimate meaningful data from huge data set available in distributed environment. In this research article we discussing about challenges and issues in analysis process of huge data sets belongs to distributed environment along with various analytical tools and their methods which are used to analysis for Big Data set.

Case Study of Analysis Techniques

Data are the pure entity which are necessary and most meaningful resource in this age of communication and globalization. There are various methods, many tools, numerous techniques and thousands of applications thatare feasible for analysis of huge data set in distributed environment [9]. This section focuses on obtaining results of analysis of classification and clustering methods which are valid across big data set available in distributed environments. Classification and Clustering encompasses wide variety of analytical techniques, methods and algorithms for the analysis of Big data. In following various tools and techniques ofclassification and clustering algorithms along with their capabilities, major findings and drawbacks based on previous researches has been defined.

Following are the basic classification techniques which are used in various data mining tasks.

- **Decision Tree Classification:** This is the technique which is in the form of a tree structure, it breaks the large scale data sets into smaller and smaller subsets and at the same time decision tree is incrementally developed. Decision trees are able to handle both categorical and numerical data and this is also used for the problem classification and regression, and for the prediction. In this technique decision tree classifier is a class capable of performing multiclass classification and it is a non-parametric supervised learning method used for classification and regression. Decision tree technique defines the path from leaf to the root and every branch of the tree stands for an outcome for the attributes and this technique boost the prediction model with accuracy, reliability, stability, portability and ease of interpretation.

  Features and Drawbacks: Main feature of decision tree algorithm is that it requires less effort for data prediction during preprocessing and it does not require normalization of data. The other advantage of decision tree classification that it does not require scaling of data as well and missing values of data set does not affect the process of building of tree structure. The main problem with decision tree classification is its instability and inaccuracy in the process of prediction and the other drawback of this technique is that it is very time consuming while training and if there are any small changes in data set can cause large modification in structure and its calculation.

- **Bayesian Classifiers:** Bayesian classification, is a statistical classifier and can predict class membership probabilities. On this technique, prediction of a member from any class can be classified such as the probability that a given tuples belongs to a particular class. In this classification Bayes theorem is used to predict the occurrence of any event which is based on level of belief, expressed as probability. Following is the equation used in this classification:

  $$P\ (c|x) = P(x|c)\ P(c)\ /\ P(x)$$

  $$P(c|x) = P(x1\mid c)\ x\ P(x2\mid c)\ x\ \dots\ P(xn\mid c)\ x\ P(c)$$

  Where, P (c|x) is the posterior probability according to the predictor (x) for the class(c). P(c) is the prior probability of the class, P(x) is the prior probability of the predictor, and P(x|c) is the probability of the predictor for the particular class(c).

  Features and Drawbacks: This technique come with features pack qualities, it is quite faster than other classification method and all features of this technique contribute equally so it becomes fast at the same time and saves the efforts. Another advantage of this technique is that it is more suitable for multiclass prediction problem and it requires less training data. The Bayesian classification technique is better for categorical input variables than numerical variables. Apart from all the capable features, its all-predictors features are independent, which is not suitable for real life problems and that's why it is not applicable for all the use cases of real-world scenario.

- **Neural Network Classification:** Neural networks are more of a complex model, which mimic the way the human brain develops classification rules. A neural net consists of many different layers of neurons, with each layer receiving inputs from previous layers, and passing outputs to further layers. It is used for feature categorization which are very similar to fault-diagnosis networks, except that they only allow one output response for any input pattern, instead of allowing multiple faults to occur fora given set of operating conditions. The ability of neural networks to detect and assimilate relationships between a large number of variables is becoming increasingly relevant to businesses that wish to effectively mine and understand huge data set term as Big Data in distributed environment.

  Features and Drawbacks: Neural network has the ability to provide the data to be processed in parallel, which means they can handle multiple tasks at the same time and influences the performance of neural network. Neural network stores the information on networks which provides the accessibility from anywhere and at any time so generation of result can be possible at any time and this one gradually being broken down which means that neural network classification techniques are reliable. Meanwhile all its advantages this technique is very costly in implementation because it uses parallel processing so many processors are required for this and that indicates its dependency on hardware's which cost all if hardware failure occurs. Neural network classification technique handles the numerical data which is very difficult to understand so the problem statement becomes more complex which makes it very difficult to work.

- **K-Nearest Neighbor (KNN):** It is simple, supervised technique used to solve classification and regression problem. It is easy to understand and implement which uses some or all the patterns available in the training set to classify a test pattern. KNN uses data with several classes to predict the classification of the new sample point. It is non-parametric technique since it doesn't make any assumptions on the data being studied,i.e., the model is distributed from the data. KNN makes predictions using the similarity between an input sample and each training instance.

  Features and Drawbacks: KNN technique is quick in calculation and simple in implementation. While KNN is very versatile, accurate simple algorithm to interpret, most of the time it does not need to compare with another supervised model. Easy in implementation is the main advantage, but it also has many disadvantages like KNN doesn't use the training data points to make any generalization which makes it significantly slow as the size of the data in use grows.
- **Support Vector Classification:** It is a supervised machine learning algorithm that can be used for both classification or regression challenges. In this each data item has been plotted as a point in multidimensional space with the value of each feature being the value of a particular coordinate and after that classification by finding the hyper-plane that differentiates the two classes very well took place.

  Features and Drawbacks: SVM classification works very fine with clear margin of data wangle working with distributed data set and it is very effective inhigh dimensional spaces where dimensional are greater than number of samples. The key factor of this classification technique is that it aims to finalthe hyper plane that separates He data points in the training set with farthest distance. Having plenty of features it seems the best one but when it comes to the large data orhuge data sets it does not perform well because required fine to training is too higher and because of developing of target classes performance of SVM classification techniquegoes down because it does not able to handle the data sets having more noise.
- **Linear Regression:** It is a simple supervised regression classification technique that is used to predict the value of a dependent variable for a given value of the independent variable by effectively modeling a linear relationship between the input and output variables using the given dataset. In this independent and dependent variable are related linearly. In aspects of data analysis and data mining this technique is used of regression rather than classification.

  Features and Drawbacks: Yet linear regressions are simple technique and provides sati's factory result with reliability but it is not suitable for classification because it deals with continuous values whereas classification problems mandate discrete values.

- **Logistic Regression:** It is a classification method used to find the probability of event success and event failure. This technique is widely used in event driven fields to predict the possible solution of the problem and it is used when dependent variable is binary (0,1) means it is in the form of true or false. This classification technique is a linear one for the given data set and then transform in non-linearity in the form of sigmoid function. It is based on sigmoid function where output is probability and input can be from -infinity to +infinity. Let's discuss some advantages and disadvantages of Linear Regression.

  Features and Drawbacks: This is the technique very easy to inclement interpret, efficient reliable and easy to train and it makes no assumption about distribution ofclasses in feature space. In this technique prediction of classes are natural probabilistic and this can equity extendable to multiple classes. The biggest disadvantage of this technique the problem of non-linear data set can't be solved because it has linear decision surface and linear separable data is rarely found in real word scenario.
- **Association Classification Rule:** This classification rule is generated by CBA-RG which is known as classification association rules (CARs), as they have a predefined class label or target. Classification using association rules combines association rule mining and classification, and is therefore concerned with finding rules that accurately predict a single target class variable. As mention it is the combination of two important anddifferent fields aims to develop accurate and interpretable classifiers by using association rules. The main objective of this is to find all the rules in data that satisfied the specified minimum confidence which is very useful and reliable for classification of huge data set in distributed environment.

  Features and Drawbacks: This is reliable technique for classification, focuses on data which is useful and informative. This technique reduces the error and improves the performance of data set. Association rule classification is very useful to analysis of huge data set and it also very reliable for prediction of customer behavior. The main advantage of this classification technique is that it is for programmers, they can easily use this rule to build programs which are capable to cope up with machine learning. While having feature rich qualities it also has drawbacks and disadvantages as its biggest drawback is that this technique is very exhaustive search-based classification technique. The major problem with this classification technique that it does not scale well when huge data set are in the pipeline so this will create many problems and takes lot of effort to handle big data set in distributed environment.

Clusters are nothing but the grouping of data points such that the distance between the data points within the clusters is minimal. Following are the basic clustering techniques which are used in analysis of huge data or Bigdata set available in distributed environment.

- **Partitioning Clustering Methods:** In this clustering technique all objects are considered initially as a single cluster. The objects are divided into no of partitions by iteratively locating the points between the partitions. This clustering technique decomposes the data sets into a set of disjoint clusters and it is one of the most popular choice for analyst to create clusters for analysis process. Clusters are creator in the basis of their characteristics of data point and there is a need to specify the number of clusters for this clustering method. K-means clustering PAM clustering CLARA clustering etc. are the famous clustering method which are based in pardoning clustering technique.

Features and Drawbacks: This clustering technique is very simple and easy to implement for the large-scale data set. Partitioning clustering method guarantees full convergence and provides a warm start for the positions of centurions. The biggest feature of this is that it easily adapts to new examples and can generalize to clusters of different shapes and sizes. Having many capabilities, it also has a major problem is that itis being dependent on initial values and it takes lot of effort with recitative processes to get the better result. Partitioning techniques for clustering have trouble in analysis of data where clusters are of varying sizes and density.

- **Hierarchical Clustering Methods:** This clustering technique creates the clusters on the basis of distance metrics and it implies on two approaches the first one is agglomerative also known as bottom-up approach which considers every data point as a starter in its singleton cluster and the two nearest clusters are combined in each iteration until the two different points belong to a similar cluster and the second one is divisive method also known as top-down approach which performs recursive top-down splitting.

  Features and Drawbacks: Hierarchical clustering is best for small data sets and this technique sums up the data. It is very easy to implement as compare to other clustering technique and gives the best results in some cases where data sets are small. Thisclustering technique is fast accurate and reliable for small data set but when it comes to the large scale of data it does not work well and it became very difficult to handle different size cluster and convex shapes.
- **Density Based Clustering:** In Density-based clustering methods data objects are divided into core points, border points and noise points. All the core points are connected together based on the densities to form cluster. In this type of clustering, clusters are created in the basis of density of the data points which are defined in data space. Density-based clustering approaches are much better than other clustering techniques because this has some special features like clustering arbitrary shape groups of data regardless of the geometry and distribution of data, robustness to outliers, independence from the initial start point of the algorithm, and its deterministic and consistent results in the repeat of thesimilar algorithm.

  Features and Drawbacks: The main advantage of this technique is that it does not esquire a priori. Specification of number of clusters and it this technique is able toidentify noise data while clustering. Density clustering is able to signal size of the arbitrarily and the shape of a clusters. The main drawback of this technique is that it is not effective work precedes reliable result when there are variations in density clusters and dimensions are high.
- **Grid Based Clustering:** In this clustering data is represented into the form of gridstructure sometimes also known cell structure. In this technique value space, which is surrounded by data points rather than the data points themselves. Grid based algorithm partitions the

data set into no number of cells to form a grid structure. To form clusters, Grid algorithm uses subspace and hierarchical clustering techniques and after partitioning the data sets into cells, it computes the density of the cells which helps in identifying the clusters.

Features and Drawbacks: One of the greatest advantages of these algorithms is its reduction in computational complexity. This makes it appropriate for dealing with humongous data sets. The biggest drawback of this technique is that it needs a large number of parameters. This technique is more efficient and reliable but more costly and time consuming.

- **Model Based Clustering:** This technique is for optimization the fir between the data and some mathematical models. It is a clustering method in which data points sets are connected together based on various strategies. There are two approaches for model-based algorithms one is neural network approach and another one is statistical approach. In the model- based approach, parameter estimation becomes difficultwhen there are too few data points in each cluster. As a result, the BIC scores of some ofthe models are not available when the number of clusters is large.

  Features and Drawbacks: Model based clustering techniques helps the applications of cluster analysis to formulate the probabilistic model and the cluster shapes aimed for more explicit. This technique isreliable accurate and potable for small scale of data set but for large scale data set itsperformance is very poor and time consuming. In this technique estimation of parameter is verydifficult when there are too few data points in each cluster.

Summary of Case Study

Various methods, tools and techniques are in use for analysis of big data and there are plenty of work done in past to perform big data analysis in distributed environment. In every single seconds internet users are increasing, peoples are connecting globally for their business purpose through along with their customer via communication and networking and this is the reason that millions of data from many different fields are generated enormously in distributed environment. Analysis of such massive sets of data and finds valuable group or set of data from distributed environment become very critical and challenging issue for all of us. After the analysis it is crystal clear that the different tools are focused on individualtask such as real time working operations or some are focused on batch processing etc. For the analysis different techniques were used like cloud computing, machine learning, quantum computing data stream processing and intelligent analysis along with classification and clustering method. Above clustering and classification techniques are very well known by the analyst but there is still a need to come up with a tool which carry almost all kinds of properties and features of above tools and able to clear all kinds of drawbacks and disadvantages in aspects of Bigdata analysis, by this ability to predict and bring out meaningful data set will be more accurate, reliable and helpful.

CONCLUSION

In this age of communication technology everything is related to data processing in distributed environment in which all kinds of task took place through online mode and all the stake are connected globally through internet. This case study presents the analytical part of various clustering and classification techniques and models which are used for the mining of big data. In future there is need to design such tools which can work in multiple environments and perform different task at same the time effectively, efficiently, accurately and with reliability.

References

1. Elgendy, N., and Elragal, A. (2014). Big Data Analytics: A Literature Review Paper. The 14th Industrial Conference on Data Mining (ICDM), Petersburg: Springer-LNCS.
2. Chen, H., Chiang, R., &Storey, V. (2012). Business Intelligence and Analytics: from Big Data to Big Impact. MIS Quarterly, 36 (4), 1165-1188.
3. McAfee, A., and Brynjolfsson, E. (2012, October). Big Data: The Management Revolution. HBR, 3-9.
4. K Murali Gopal, Dr. Pragnyaban Mishra, and Dr. R. P. Singh, “BDADE: Challenges, Research Tools and Literature in Big Data Analysis in Distributed Environment”, International Journal of Scientific & Technology Research Volume 9, Issue 03, March 2020, pp. 748-753.
5. Umasri. M.L, Shyamalagowri. D , and Suresh Kumar. S ,Mining Big Data:- Current status and forecast to the future Volume 4, Issue 1, January 2014 ISSN: 2277 128X.
6. Ravi Ranjan and Aditi Sharma ―Evaluation of Frequent item set Mining Platforms using Apriori and FP Growth Algorithm‖. 4th International Conference on Computers and Management.
7. https://www.getsmarter.com/blog/career-advice/big-data-analysis-techniques/**.**
8. https://data-flair.training/blogs/best-big-data-analytics-tools/.
9. https://bigdata-madesimple.com/top-30-big-data-tools-data-analysis/.