IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 5, May 2016*

# Health-Care Proposal Tool using Big Data Techniques

**Utkarsha B[1], Priyanka C[2], Dishari S[3], Aneri P[4]**

**Abstract:** A decision-tree is a structure which can represent any algorithm in a tree or even graph with the nodes and branches with some associated outcome in terms of weight and probability. Basically here, we are using large datasets gathered in clinical records related to a patient's health care issues and their medical reports to predict his or her future health diagnosis and accordingly recommend the required medication. The recommender tool being implemented in Hadoop framework uses popular classification algorithm named as C4.5. We are trying to improve the performance of the decision-tree algorithm utilizing the appropriate Map-reduce model and forward it to Hadoop framework in addition to bagging method additionally with random-subspaces in ensemble classification in order to improve efficiency and scalability.

**Keywords:** C4.5, algorithm, decision-tree, ensemble classification, Hadoop framework, Map-reduce, Recommender tool, PHR, HRS.

## I. INTRODUCTION

Recommender tool (Rt) is a system which provides its clients the most appropriate data through an information system or e-business system and has a great advancement in recent years. A most popular example is the Amazon's service suggesting service for items. A trust for the thought behind recommender tools is adapted slowly in the present generation to the unique necessities of the health domain.

In the recent era large volume of data has been gathered in the clinical databases showing the patient's wellbeing states, for E.g. therapeutic reports. Henceforth, digital data i.e. the computerized information easily accessible to the patient's decision making has increased numerously but still is spread all over the sites. Also the PHRS are helping to gather an individual's medical history details and grant access to the owner in addition to some approved health-care experts so that they can help that individual anywhere and time to figure out the ailments and the required diagnosis accurately. As per a recent survey the adults use internet and maximum percentage of them have the complete online study and accordingly act for data in regards to health to ailments, analyse and diverse treatments. This gives a rise to a stronger patient-doctor relationship as educated patients bring up issues or talk about the various choices of treatment available and also suitable.

The main purpose of such system is to provide its clients the proper medical data which is intended to be profoundly unique to the therapeutic development of the patient connected to that PHR. The health experts who take a shot with the given PHR are prescribed with related medicinal data. In addition to that it might also be prescribed to laymen studying their own PHR. Contingent upon a client's therapeutic ability a HRS ought to recommend medical knowledge, which is comprehendible to that client. In this act a new idea is to make use of various machine learning based information characterisation methods over the cloud in order to tackle a issue of real time classification. Initially deploying and testing the classification algorithm that was created on the cloud would be the actual ultimatum. This paper comprises of not only the learning of the distinctive techniques of machine learning and its applications, but also includes recognizing the algorithm and nature which will be most suitable for this specific classification issue. Thus building a proper decision-tree with the help of given dataset is referred to as the decision-tree algorithms, as such an inducer algorithm. Ordinarily the objective here is to minimize the generalization error while locating the ideal decision-tree.

Decision-tree algorithms poses different appealing characteristics: straightforwardness, ability to understand, simplicity, non-parametric, capability to handle mixed sort of data. The study shows that a decision-tree is affected from arrangement of named preparing cases spoken to by a tuple of attribute values and a class label. Due to these properties it has a standout amongst the various learning algorithms. Also the decision-tree learning is a greedy approach, top-down parsing and a recursive procedure beginning with an empty tree and complete training data set.

Basically most inducers perform just the developing phase. Inducers can be top-down or bottom-up in approach. For example, ID3, C4.5 , CART etc. having two stages called as developing and pruning.

Use of growing and pruning is done in the activity for the top-down algorithm of decision tree. The algorithms are basically greedy by nature and build the decision tree in a recursive way. In every emphasis, the algorithm considers a segment of the training set utilizing the result of a discrete capacity of the input properties. The finalizing of the most appropriate function is made with respect to some measures. After the choice of a proper split, every end system further subdivides the training set into further littler subsets, until no split increases adequate part measure or a halting criteria is fulfilled.

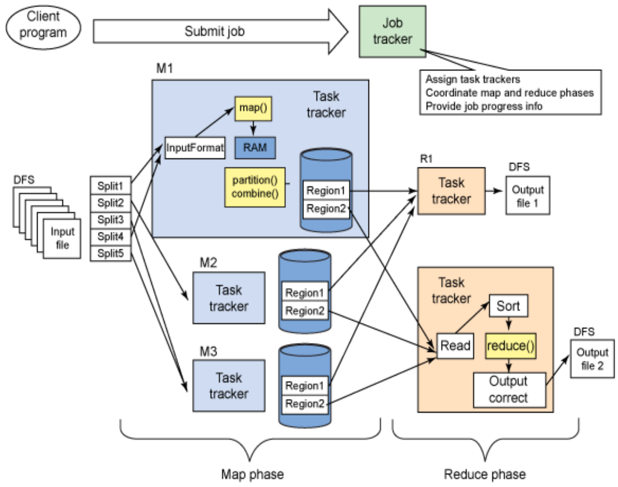
DOI 10.17148/IJARCCE.2016.55157

## II. RELATED WORK

The data-mining algorithms are implemented using various tools in accordance to their requirement. Ex. Weka, Mat lab, Hadoop, Map-reduce etc. Let's have a small glimpse over these tools and their implementations with respect to pros and cons respectively.

### A. Hadoop

Hadoop is actually a programming architecture for creating applications that quickly process large quantity of information in parallel on numerous groups of nodes. It is basically a piece of venture supported by the Apache s/w Foundation.
Hadoop [4] makes use of java language for its operations. Initially was considered on the premise of Google's Map-reduce where one complete application is segregated into small chunks. It can provide tremendously whatever power is required for the application and versatility choice to a dispersed framework. The library present inbuilt based on programming identifies and handles failures at the application layer of the OSI protocol model. Hadoop provides reasonable and dependable capacity. It computes parallel processing over a vast group of machines all executing simultaneously. Also it has the ability to convey an exceptionally accessible administration on top of a group of PCs, each of which might be inclined to failures.
Map-reduce programming paradigm[4] which is a nonexclusive execution model is also effectively and necessarily deal done with Hadoop. Map-reduce is a circulated programming paradigm expected for substantial group of frameworks that can work in parallel on an expansive dataset.
The following figure depicts the same as:



Every process in here in working in co-ordination and totally parallel way. Initially, dataset being partitioned into various chunks and then processed parallel over various mappings undergone over every chunk simultaneously. Mapping is a method in which every subset of the dataset is mapped over some attribute value using some appropriate functions E.g. Hashing-function applied over the file systems. The Reduce structure sorts the yields of the maps, which are then given as data to the lessen undertakings. Both the data and result of the occupation are put away in the file system. Because of parallel processing nature of Map-reduce, parallelizing information mining algorithms utilizing the Map-reduce model has got a noteworthy consideration from the examination group subsequent to the presentation of the model by Google.

### B. Classification Algorithms

A] Naïve Bayes[4] is a classifier, used popularly for text categorization.it is based on implementing the Bayes theorem. It legitimately fits in accordance to Map-reduce engineering. The processing time has been greatly mitigated with the Naïve Bayes implementation. Has a great execution but still some further modifications can be made in order to bolster block key worth updating mechanism.

B] Gong-Qing[3] Wu actualized C4.5 decision tree characterization calculation on apache hadoop. It is a statistical classifier. Various improvements were needed to this basic decision tree algorithm for effective performance. In this scenario, while developing the stowing troupe based decrease to build the last classifier numerous copies were found. These duplicate copies couldn't have maintained a strategic distance from if appropriate information transmission and partitioning strategy have been connected.

C] Zganquan[9] sun investigated the materialness of SVM on Map-reduce stage. It analyzes the data for classification and regression. It is not a probabilistic classifier. SVM is a linear binary classifier.

Zganquan[9], throughout his studies he found that Map diminish can reduce the processing time and calculation time. The dividing strategy was exceptionally indistinct. Support vector machines have been utilized effectively as a part of numerous order errands. Their calculation and capacity need increment quickly with respect to the processing vectors. No relationship between the dividing strategy and the execution could be inferred. Formation of a global classifier used all over which is optimally reduced in a definite iteration from continuous recursive iterations is seen.

D] Some non-parametric algorithms like kNN act the hero, in scenarios where the important part of the actual information lies over the hypothetical made. kNN is likewise an apathetic calculation this suggests it doesn't utilize the preparation information focuses to do any speculation. Along these lines, the preparation stage is lovely quick. Whenever speculation is not present in the lines it implies that the algorithm knows the complete processing information. kNN settles on choice in light of the whole preparing information set.

To further demonstrate a logical comparison between contrasted Map-reduce kNN with respect to consecutive kNN , which resulted that Map-reduce kNN out performs the successive kNN with datasets of Larger Size.

DOI 10.17148/IJARCCE.2016.55157

Hadoop being stage free gives the client the adaptability to have Heterogeneous framework to be interconnected. Thus the client can utilize effectively existing equipment to setup a Hadoop group.

## III.PROPOSED WORK

The C4.5 algorithm is worked out under Map-Reduce technique with some enhancements. Fig 1. Demonstrates the improved architecture of C4.5 algorithm. Architecture shown below basically has three stages i.e. Segmentation, Mapping and Reduction.

### A. Stage( i) Segmentation:

In this stage, smaller segments are formed, which is further evaluated we divide a large data LD into smaller segments or parts (LD1, LD2, …..LDn) using bootstrap sampling methodology.

### B. Stage( ii) Mapping:

Mapping is basically correlating the dataset with respect to some functions or attributes in order to segregate the most common datasets.

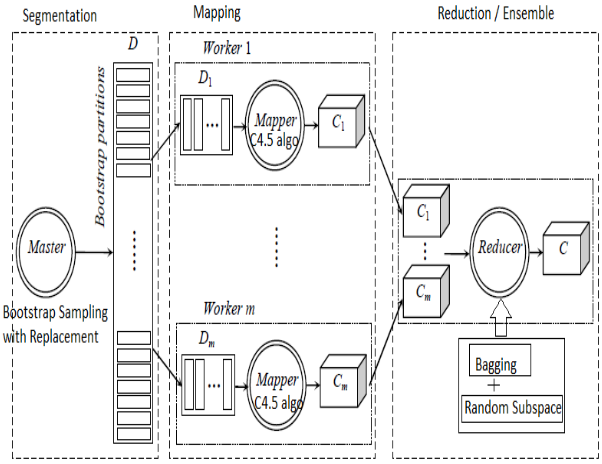
Each mapper class is responsible for constructing a base classifier. Key1 is unique identifier for document name connected with data set D1. Value1 is a derivated content through the data set created using bootstrap sampling with replacement. Map constructs a classifier c1 from value1 using algorithm known as C4.5 and middle results are presented to the Map-reduce paradigm. For submitting the middle results emit function is used.

### C. Stage( iii) Reduction

Key1 is the identifier for document name related the dataset D1. Value1_list is the blend of intermediated results presented by Mapper class. Reduction operation constructs base classifiers from each esteem in value1_list, and after that as per bagging, reduction operation builds a classifier c1, and results acknowledged to Map-reduce model using function known as emit.

Here we are recommending the idea of blending bagging methodology along with random-subspace to get improved performance in reduction stage of C4.5 algorithm in map reduce framework. Generally, Bagging works by training data set and producing irregular autonomous bootstrap samples i.e. random samples. Base classifier qualified with making use of samples.

Random-subspace strategy uses the method of learning base classifiers from data feature-space from random subspaces. It arbitrarily chooses a part from features toward the begin, uses a deterministic form of the base level algorithm. In this consolidated strategy, datasets are modified in dual ways. In the first place, the alteration is performed in the data set by taking bootstrap recreates bSi = bXi1,bXi2,...,bXin of the preparation set bS = (bX1,bX2,...,bXn). After that, an alteration is performed in the element space on each bootstrap recreate taken from the dataset.



## IV. CONCLUSIONS

In this document, we are specifying a Recommender tool based on health-care issues. We have discussed implementation of various algorithms related to data-mining, specifically the classification algorithm C4.5 to be implemented in Hadoop Framework. We have additionally used the Map-reduce Ensemble classification method for making reduction in the bias of variance between bootstrap values and real entity values to increase the performance of C4.5, scalability and efficiency. Finally, focus of this proposal is to make improvement in the performance of C4.5 algorithm in Map-reduce framework by doing some advancement in bagging method in reduction step of algorithm.

## ACKNOWLEDGMENT

We would want to express our genuine appreciation to our guide **Prof Priya V.** SITE SCHOOL for supporting continuously for our exploration paper work. We are obliged for his ceaseless backing and grateful for motivating us to concentrate more over this subject. We are extremely happy to experience such fascinating method for research paper composing.

## REFERENCES

[1] "Bootstrap Bias Corrections for Ensemble Methods" G Hooker, L Mentch , 2015

[2] "A Personalized Framework for Health Care Recommendation" SB Ahire, HK Khanuja - Computing Communication Control, ieeexplore.ieee.org, 2015

[3] Gong-Qing Wu, "MReC4.5: C4.5 Ensemble Classification with MapReduce", chinagrid 2009

[4] "A Survey on Data Mining Algorithms on Apache Hadoop Platform" D Nandakumar, N Yambem - International Journal of Emerging, 2014

[5] "Parallel random prism: a computationally efficient ensemble learner for classification.", F Stahl, D May, M Bramer , Springer,2012

[6] "A mapreduce implementation of C4. 5 decision tree algorithm" W Dai, W Ji, International Journal of Database Theory , 2014.

[7] "Online bagging and boosting. In Artificial Intelligence and Statistics", Morgan Kaufmann, 2001.

[8] H. Li and X.M. Hu,"Analysis and Comparison between ID3 Algorithm and C4. 5 Algorithm in Decision Tree", Water Resources and Power, 2008.

[9] Zhanquan Sun,"Study on Parallel SVM Based on MapReduce", in conference on worldcomp2012

DOI 10.17148/IJARCCE.2016.55157