ISSN: 0974-2115
www.jchps.com Journal of Chemical and Pharmaceutical Sciences

Analysis of thyroid syndrome using K-MEANS clustering algorithm

B.Jothi*, S.Krishnaveni, Jeyasudha.J
Department of software engineering, SRMU, Chennai, Tamilnadu, India
***Corresponding author: E-Mail: jothi.b@ktr.srmuniv.ac.in**

ABSTRACT

In the field of medical research, massive amount of data are generated from different medical sources like health care, PHI and sensor devices which makes data become available for decision makers and research experts to collect and analyze knowledge about patient health. Introducing big data analytics in health care serves as a good tool for cleansing, transforming and retrieving valuable information from large medical data sets. Big data is a collection of massive data sets with volume, velocity and veracity which is more critical to handle. In this paper K-MEANS clustering algorithm is used for training and analyzing thyroid related data sets to say whether the patient suffers hypo or hyper thyroidism symptoms.

KEY WORDS: Big data, data mining, analytics, decision making, TSH- thyroid stimulating hormone.

1. INTRODUCTION

Medical data challenges and strengthens mass collaboration with new techniques and cost driven methods to be implemented to benefit patients. Research across all most all medical organizations are using it to develop new products and services, and also monitor them by how people extract a valued information from large medical datasets to analyze their health conditions.

Thyroid function test, looks for the levels of thyroid-stimulating hormone (TSH) and thyroxine (T4) in the blood to check whether the patient suffers from hypo or hyper thyroidism. If the values of TSH is high and T4 is low ,the patient suffers hyper thyroidism and if TSH is low and T4 is high the patient suffers from hypo thyroidism and yet, most of the big data analytics struggle to measure the true value from the large medical sets in order to deliver insights about the patient who require serious and correct medications To find the cause of an underactive thyroid gland hypothyroidism or pituitary gland or the hypothalamus, TSH levels helps to determine the causes of problems associated with damaged thyroid gland or some other cause such as a problem with the pituitary gland or the hypothalamus.

To leverage medical data, from huge volume of data set, we are passing it into a hadoop framework to cleanse and extract knowledge from it. which makes the medical data sets to be trained using suitable mining algorithm ,so data's are fed as input over cluster of nodes in network by HDFS and which in turn fed as input and output to mapper and reducer to generate the collect output after analysis.

Hadoop: Hadoop is java program implemented by apache software foundation for processing huge scalable data which comes with different velocity and variety .Hadoop is a kind of storage platform designed to process very large data sets by using parallelism mechanism across hundreds to thousands of computing nodes. Hadoop provides a cost effective storage solution for processing large data volumes with no required format requirements. Hadoop has two main components- HDFS and YARN.

HDFS: Hadoop distributed file system stores metadata and also data related to different domain application separately by using Name node and Data node blocks. Where the Name node stores all the roots and directories of metadata and the Data Nodes stores information about collection and retrieval of large data sets .TCP protocols are used in communication among cluster of nodes.

Name Node: Name node stores hierarchy about all files and directories which is represented as in node record attributes. This record attribute contains permission, modifications to access time and disk spaces related to I/O processing by dividing large data among 128 megabytes, but user selectable file-by-file, and each block of the file available is independently replicated at multiple Data Nodes.

Data Node: Each block of lookalike on a Data Node is represented by two files in the local native file system. The first file contains the data itself along with which the second file records the block's metadata including checksums for the data and the availability of the stamp. The size of the data file would equal the actual length of the block and does not require any extra amount of space to round it up to the nominal size of the blocks as in traditional file systems. Thus, if a block is half empty it needs only half of the space of the total block available on the local drive. During startup each Data Node networks with a Name Node and performs a handshake. The purpose of the handshake is to authenticate the namespace ID and the software update of the Data Node. If either does not suit with that of the Name Node, the Data Node would automatically shuts down.

HDFS Client: User applications have abilities to access file using HDFS client by exporting HDFS file system interface. This provides righties to read, write and delete files from directories using application path in Namespace to support multiple replica of same block which can be used during data loss. The application reads file from HDFS client from the Name node for the list of data nodes from the block of files. Then the list is arranged in network topologies, which distance from the client, then the client directly asks for the data node to transfer the needed block. When the first block is completed, the client request a new Data Nodes to be chosen to host replicas of the next block

a new pipeline that has been organized, and the client sends the further bytes of the file. Applicable options of Data Nodes for each block is likely to be different and stores snapshots of data to handle during error.

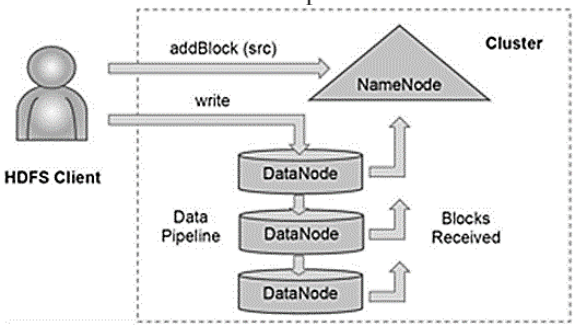

Figure.1. HDFS Client Creates a New File

Checkpoint Node: The Name Node in HDFS, in addition to its primary role serving client requests, can alternatively execute either of two other roles, either a Checkpoint Node or a Backup Node. The role is specified at the node startup.

The Checkpoint Node periodically combines the existing checkpoint and journal to create a new checkpoint and an empty journal. The Checkpoint Node usually runs on a different host from the Name Node since it has the same memory requirements as the Name Node. It downloads the current checkpoint and journal files from the Name Node, merges them locally, and returns the new checkpoint back to the Name Node.

Backup Node: A recently introduced feature of HDFS is the Backup Node. Like a Checkpoint Node, the Backup Node is capable of creating periodic checkpoints, but in addition it maintains an in-memory, up-to-date image of the file system namespace that is always synchronized with the state of the Name Node.

Upgrade and File system Snapshots: There is a possibility of file corrupting during file upgrade process due to errors, bugs or increase in human mistakes. The purpose of creating snapshots in HDFS is to minimize potential damage to the data stored in the system during upgrades. Snapshot technique in HDFS allows administrators to store the current state of file system, so if any updates results in data loss using rollback mechanism in HDFS, the files can be easily retrieved.

Map Reduce: The job of Map Reduce is to split large data in small independent chunks, which is then organized into key and value pairs for parallel processing. The use of parallel processing in Map reduce improves speed and reliability of clusters which return in more optimized output Mapper divides input data into different ranges by creating a input format, and then with the help of job tracker the tasks are distributed to worker nodes which in turn reduces the output into smaller key and value pairs. Finally the job of reducer is to collect various results from large problem set from parallel processing cluster nodes and writes the data back to HDFS .Thus, the reduce is able to collect the data from all of the maps for the keys and combine them to solve the problem. Due to the iteration of steps repeated to support parallelism map reduce alone won't serve as a flexible framework in processing large medical data sets. So a suitable machine learning algorithm have to be implemented to handle massive medical data sets like thyroid and to extract information.

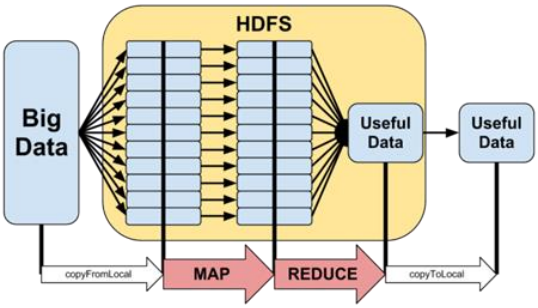

Figure.2. Map reduce

K-means on Map reduce: K-MEANS is a machine learning algorithm used for cleaning and retrieving information from huge data sets. It is a kind of clustering algorithm which works by dividing the Data points D from a large set in K cluster nodes. Thus by dividing and initiating D data points, K nearest position is calculated from the cluster of nodes and the step is repeated in a iterative way to map and reduce data in smallest key and value pairs. The pseudo code for mapper and reducer functions for k-means clustering algorithm. Basically, mappers read the data and the centroids from the disk. These mappers then assign data instances to clusters. Once every mapper has completed their operation, reducers compute the new centroids by calculating the average of data points present in each cluster. Now, these new centroids are written to the disk. These centroids are then read by the mappers for the next iteration and the entire process is repeated until the algorithm converges. This shows the disk access bottleneck of Map Reduce

for iterative tasks as the data has to be written to the disk after every iteration. Similarly terabytes of Thyroid syndrome medical data sets are fed into hadoop framework, which is processed using K-MEANS algorithm by leveling the different TSH and T4 values in to discriminate set of Key and Value pairs .And final output from hadoop framework is clustered using K-MEANS algorithm to indicate which thyroid syndrome a patient suffers from.

Algorithm:

Input: Data points D, Number of clusters K

Step1: Initialize K centroids randomly

Step 2: Associate data point D with the nearest centroids. Divide data points into K clusters

Step3: Recalculate the position of centroids

Step4: Repeat step 2 and step 3 if there no changes in membership of data points

Step5: Data points with cluster memberships

Experimental Setup: Thus by using a K-MEANS machine learning algorithm a large amount of PHI medical set related to thyroid syndrome is fed as input, to hadoop framework to analyze the pattern of disease that could occur by variations in hormone levels by categorizing class attribute to say whether the patient has hyper or hypo thyroid syndrome symptoms. The following data in table 1 shows the thyroid syndrome data taken as input to find the symptoms of hypo and hyper thyroidism.

Table.1. Thyroid syndrome data set

Class attribute	T4-resin uptake test (D data point)	Maximal absolute difference of TSH value (k centroid)	Train Key pair(D,K)	Diagnosis Test
1	5	2.7	(5,2.7)	hypo
2	6	4.2	(6,4.2)	hypo
3	0.5	3.1	(0.5,3.1)	hyper
4	13.5	1.5	(13.5,1.5)	hypo
5	12.5	0.4	(12.5,0.4)	hyper
1	5.5	3.1	(5.5,3.1)	hyper

2. CONCLUSION

K-MEANS machine learning algorithm is implemented in hadoop framework by finding the data centroid among nearest data node using map reduce framework .Similarly thyroid syndrome datasets uses K-MEANS algorithm for calculating the average of data point and writes it to the disk with new centroid calculated ,then the data are immediately passed to next iteration from mapper to reducer with sequence of iterations to extract a valuable output to predict whether the patient is suffering due to hypo-thyroidism or hyper-thyroidism syndrome .Thus the overhead in disk I/O and iterations in map reduce are overcome by using K-MEANS algorithm.

Future work: In future K-MEANS algorithm can be more optimized to increase the speed throughput of data set by adding more features to provide information like what medications can be given to particular symptoms of disease

REFERENCES

Feldman B, Martin EM, Skotnes T, Big Data in Healthcare Hype and Hope. Big Data in health Care, October, 2012, Dr. Bonnie 360.

Fernandes L, O'Connor M, Weaver V,Big data, bigger outcomes, JAHIMA, 2012, 38-42.

https://archive.ics.uci.edu/ml/datasets/Thyroid+Disease.

Lam C, Hadoop in Action. Manning Publications Co., New York, Google Scholar, 2012.

Official us doc, Thyroid Function Tests, U.S. Department of Health and Human Services National Endocrine and Metabolic Diseases Information Service, 2005.

Singh, Dilpreet, and Chandan K Reddy, A survey on platforms for big data analytics. Journal of Big Data, 2(1), 2014, 1-5.