International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-11 Issue-2, December 2021

COVID-19 Data Clustering and Testing with K-Means Mapper and Reducer

A. Anusha, K. Kishore Raju

Abstract: ***Due to the emergence of a new infectious disease (COVID-19), the worldwide data volume has been quickly increasing at a very high rate during the last two years. Due its infectious, and importance, in this paper, K-Means clustering procedure is applied on COVID data in MapReduce based distributed computing environment. The proposed system is store, process and tests the large volume of COVID-19 data. Experimental results had been proved that this process is adaptable to COVID-19 data in the formation of trusted clusters.***

Keywords**: *K-Means Clustering, MapReduce, Unsupervised Machine Learning and Covid.***

I. INTRODUCTION

Big data provides a tremendous amount of data to researchers, health personnel, and biologists, ability to make informed decisions about how to combat the COVID-19 virus. These data can be used to continuously track the infection on a global scale and to spur medical innovation. [1]. It can assist in predicting the blow of COVID-19 on a specific location and the entire population. It aids in the development and investigation of innovative handling procedures. Big data can also give people with potential sources and opportunities, assisting them in dealing with a difficult issue. In general, this technology offers data for disease transmission, migration, and health monitoring and preventative systems to be analysed. Big data analytics methods are particularly suited for measuring and justifying COVID-19's global impact. [2]. This tool is similar to Hadoop's MapReduce, and is one of the new framework core processing building pieces. We can do such concurrent computations using the MapReduce framework without thinking about issues like consistency, fault tolerance, and so on [3]. As a result, MapReduce allows programmers to build code logic without having to worry about the system's design difficulties. The software reliability model refers to the appearance of a random process that defines the behavior of software failures from time to time. Software reliability models appear when people are trying to understand the features of how and why software fails and trying to calculate softwarecredibility. There is no personal model that can be used in all situations. No model is complete or even representative. Most software models have the following components: A mathematical function that has reliability with elements. The mathematical function is usually high-order exponential or logarithmic.

Two types of modeling methods are based on examining and collecting failure data and analyzing it with statistical inference.The respite of this work is organized as follows: Introduction is discussed in Section 1. In section-2, Optimal Points and Values on Covid-19 Patient informations. Section-3 deals with the K-Means Clustering Using Hadoop's MapReduce. Section-4 deals with Cluster Validation for Covid-19.

The details of proposed work and discussion of various parameters in which affect its performance of Unsupervised Cluster Machine Learning Using Bayes Law discussed in Section-5.The Software Reliability Model is estimated on Covid-19 in Section-6. Section-7 deals with the future perspective and conclusion

II. OPTIMAL POINTS AND VALUES

A. Optimal Points and Values on Covid-19 Patient Informations

Different optimal points and values are in generally used to optimize the models.

$$Obj = \{f0(x) | \exists x \in t, fi(x) \leq 0 \quad i = 1,2,3,\dots,n, \quad hi(x) = 0, \quad i = 1,2,\dots,p\} \subseteq Rq$$

This function named the set of achievable impartial values either Positive or Negative Patients. Consider the various Covid-19 patient inputs and are represented as the set of achievable values either Positive or Negative Cases. In the set of impartial values select one value i.e., optimal value can be represented as target Covid-19 patient as t. Therefore, eliminate other attainable values by optimization process. One point x* that is said to be optimal iff feasible and $obj \subseteq f_0(x*) + K$. To optimize this, used the convex optimization process[4].

B. K-Means Clustering Using Hadoop's MapReduce

K-Means is follows the unsupervised process that follows set of rules. Unsupervised algorithms, on the other hand, make inferences from the Covid Healthcare dataset using only input vectors and do not relate to known or labelled outcomes. It's simple to use java to run the K-Means technique on a large dataset or an Excel (.xls) file.

Manuscript received on December 08, 2021.
Revised Manuscript received on December 13, 2021.
Manuscript published on December 30, 2021.
* Correspondence Author
A. Anusha*, Department of Information Technology, S.R.K.R. Engg College, China-Amiram, Bhimavaram (A.P), E-mail: anusha.abk@gmail.com
Dr. K. Kishore Raju, Assistant Professor, Department of Information Technology, S.R.K.R. Engg College, China-Amiram, Bhimavaram (A.P), E-mail: kkrsrkrit@gmail.com

When it comes to running the Covid Healthcare Datasets at the Big Data level, however, the standard process cannot be extended any further. That's how Big Data Hadoop MapReduce Processing works [4].

K-means process had been split into neighbor finding process and recalculation of centroids processes. These two steps agree on the map and reduction steps of our mapReduce policy.

$$K-MeansMap\ (p): \quad emit\ (argmini\ \|h - \mu i\|2, (h,1))$$

Where **h** is point and μ_i is the mean with key value pair (h,1)

The reduction step is the summation that is objectively attached to key value. To put it another way, if two value pairs are given for a given key, we combine them by adding each pair's matching element. Therefore, our function is

$$K-MeansReduce\ (i,[(h,p),(g,r)]): retur(i,(h+g,p+r))$$

Produces a set of k values of the mapredius form characterized by these two functions

$$\left(i, \left(\sum_{h \in p_i} h, |P_i|\right)\right) \quad (1)$$

where P_i refers to a set of points that are close together. Then, we can calculate the new paths(Centroid of sets P_i)

$$\mu_i \leftarrow \frac{1}{|P_i|} \sum_{h \in P_i} h \quad (2)$$

each device in our distribution cluster should have a collection of present paths in order to determine the distance between a point h and each device using the map function. As a result, at the end of each iteration, we must send new pathways across the cluster. [3].

C. Steps in Procedure:

1. Initially, choose K-means($\mu_1 \ldots \mu_k$) randomly from the set h.
2. K-MeansMap and K-MeansReduce procedures are applied to h.
3. MapReduce's($\mu_1 \ldots \mu_k$) new means are calculated .
4. Newly calculated means are passed to each device in the cluster.
5. STEP 2 to 4 are repeated until means are met.

In the K-Meansmap procedure need to do the complete effort of O (knd). The whole communication cost is O (nd), and the highest number of rudiments related with a key in the reduction stage is O (n). However, meanwhile our minimize function is substitutable and harmonizing, we can use **K-Means reduce** the communication cost from each expedient to O (kd). In addition, once the **K-Means map** step is complete we need to communicate with each other to transmit new paths with size O (kd) to all the exercises in the cluster. Thus, overall in each iteration of K-Means mechanism O (knd) and includes the communication cost O (kd) when using combiners, which are one and one for all communication replicas.

D. Validation of Cluster Covid-19

Clustering outcomes valuation typically performed by some kind of measure of within-cluster dimensionality in replication process[6].

Step-1: Original dataset is devided into four cluster samples **S_1, S_2, S_3 and S_4** with **66, 68, 89, 94** cases respectively.

Step-2: Centroids are calculated for four clusters on S_1, S_2, S_3, S_4 with MapReduce and those are shown in Table 1.

Table-1

	Clusters			
	1	**2**	**3**	**4**
Factor-1 (+)	1.24	1.27	2.29	2.45
Factor-2 (-)	-0.31	-0.33	0.51	0.61

Step-3: Assign the data from the S2 to the centroids that are closest to them. The distance between the patterns in the S2, and the centroids identified previously on S1 is calculated. The nearest centroid is assigned to each S2 pattern. Allow the previously calculated centroids to be saved, making them available for classification alone in this phase, and vice versa. The actual data is shown in Table-2.

Table-2

	Clusters			
	1	**2**	**3**	**4**
Factor-1 (+)	1.25	1.30	2.30	2.47
Factor-2 (-)	- 0.35	-0.35	-0.55	-0.65

E. Unsupervised Cluster Machine Learning Using Bayes Law

$$P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{P(A|C)P(C)}{P(A)}$$

$$C\ is\ the\ class\ label\ i.e., C \in \{c1, c2, \ldots, cn\}$$

A is the observed object characteristics $A \in \{a1, a2 \ldots am\}$

$P(C|A)$ is the probability of C given A is observed called the Conditional Probability. The Conditional Probability that is C is true given that A is true, symbolized $P(C|A)$, times the probability of A is the same as the conditional probability that A is true given that C is true, denoted $P(A|C)$, times the probability of C. Both of these terms are equal to $P(A \wedge C)$ that is probability A and C are instantaneously true. If we divide all three terms by P(A) then we get the form shown. The reason that Bayes Law is important is that we may not know $P(C|A)$, but we do know $P(A|C)$ and P(C) for each possible value of C from the training data.

Assume that a disease can occur and that a test has been created to detect it. Knowing the following probability is helpful.

$$P(C) = 0.05, P(\neg C) = 0.95, P(A|C) = 0.95\ and\ P(A|\neg C) = = 0.1$$

Here P is the probability, C is the disease , A is the positive test and ¬ Not. The test result is calculated using $P(C|A)$ which is a Bayes law and

$$P(C|A) = \frac{P(A|C) \times P(A)}{P(A)}$$

$$P(A) = P(C) * P(A|C) + P(\neg C) * P(A|\neg C)$$

$$= 0.05 * 0.95 + 0.95 * 0.1 = 0.1425$$

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$P(C|A) = \frac{0.95 * 0.5}{0.1425} = \frac{1}{3}$$

By observing this value, we conclude that only one third of patients having the positive, which is very negligibly small and it is not acceptable.

10000 patients data for finding testing positive with naïve bayes is shown in Table 3

Table-3

	Disease	¬ Disease	Total
Test +	10000 × (0.05 × 0.95)	10000 × (0.95 × 0.1)	1425
Test -	10000 × (0.05 × 0.05)	10000 × (0.95 × 0.9)	8575

If you test positive, your chances of having the disease are calculated by dividing the total number of positive tests by the number of positive tests with the disease present. i.e., **475/1425** = **1/3**.

III. SOFTWARE RELIABILITY MODEL

Reliability enhancement model is a numerical model of software reliability that reflects how software reliability improves over time as errors are detected and repaired. These models help the manager in deciding how much effort to put into the test. The objective of the researcher is to test and debug the system until the required level of reliability is reached. Software Reliability refers to the capacity of a medical application to perform correctly in a certain environment and for a set amount of time. The probability of failure is computed using the formula below by evaluating a sample of all available input states [7].

$$Probability = \frac{No.of\ failed\ circumstances}{Total\ No.of\ circumstances\ under\ consideration}$$

The input space is the collection of all potential input states. To determine the authenticity of software, the author must locate the supplied input space and output position from the medical applications.

IV. CONCLUSION

This paper is discussed the vector optimization problematic through optimal points and values are considered. The set of objective values of feasible points are reflected .K-Meansmap procedure need to do the complete effort of . The whole communication cost is O (nd), and the highest number of rudiments related with a key in the reduction stage is .The replication analysis is exemplified the Healthcare application to the MapReduce based K-Means cluster solution of the Covid-19 data derived. Exemplified with rigid numbers filled the following test/disease matrix for a population of 10000 patients. Software Reliability is the potential for medical application to function properly in a specific environment and for a period

REFERENCES

1. Xia W, Sanyi T, Yong C, Xiaomei F, Yanni X, Zongben X: When will be the resumption of work in Wuhan and its surrounding areas during COVID-19 epidemic?, A data - driven network modelling analysis. Scientia Sinica Mathematica, 2020, DOI: 10.1360/SSM-2020-003.
2. Haleem A, Javaid M, Vaishya R: Effects of COVID 19 Pandemic in daily life, Current Medicine Research and Practice, 2020, DOI: 10.1016/j.cmrp.2020.03.011.
3. Stephen Boyd, Convex Optimization, Cambridge Books Publication, 2004.
4. O. Y. Al-Jarrah, P.d. Yoo, S. Muhaidat, G. K. Karagiannidis and K. Taha, "Efficient Machine Learning for Big Data: A Review" in Journal of Big Data Research, Elsevier, 2015.
5. Kollios G, Gunopulos D, Koudas N, Berchtold S. Efficient biased sampling for approximate clustering and outlier detection in large data sets. IEEE Trans Knowl Data Eng. 2003;15(5):1170–87.
6. Weijia Lu: Improved K-Means Clustering Algorithm for Big Data Mining under Hadoop Parallel Framework, Journal of Grid Computing volume 18, pages239–250(2020), Published Date:20 December 2019.
7. IEEE Recommended Practice on Software Reliability, IEEE, DOI:10.1109/ieeestd.2017.7827907,ISBN 978-1-5044-3648-9

AUTHORS PROFILE

A. Anusha, has did bachelors degree of engineering in Electronics & communication from Sri Vishnu Engineering College for Women, Bhimavaram, Currently, she pursuing her Masters in Information Technology at S.KRK Engineering College Bhimavaram. She more passionate on Network Security, embedded systems and Big data..

Dr. K. Kishore Raju, has did masters from S.KRK Engineering College Bhimavaram and Ph.D from JNTU, Kakinada..