TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 4303-4308

Enhanced Classification and DNA Security on Large Scale Dataset

Rakshitha M[1], Ranjitha V[2], Samarpita Maitra[3], Sanjeevini Nasi[4], K Amuthabala[5]

[1,2,3,4,5]School of C&IT, REVA University, Bangalore, India
[1]rakshitham38@gmail.com, [2]ranjithavammu1998@gmail.com, [3]sam29maitra@gmail.com, [4]sanjeevini2628@gmail.com, [5]amuthabala.p@reva.edu.in

Article Info
Volume 83
Page Number: 4303-4308
Publication Issue:
May - June 2020

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *12 May 2020*

Abstract

In today's rapidly growing world, handling big data is difficult as it raises many concerns like analyzing the data, ease of accessibility, efficient storage, effective management as well as enhanced security. Therefore there is a need of a mechanism that can make the machine effectively learn the way to handle data along with proper authentication of data. So in this paper, we propose a mechanism in which all the above concerns are achieved. The mechanism consists of Pre-Processing of raw data that is fed as an input for Enhanced Classification which in turn produces blocks of data with the help of Enhanced Map-Reduce algorithm that is deployed in the public cloud with the unique identification of hash code generated for every block. The authentication of data for encryption and decryption processes is carried out with the help of DNA algorithm. Thus this scheme projects a beneficial approach in terms of security analysis and numerical analysis over bigdata.

Keywords**:** *KNN classification, Hash Code generation, MD5 algorithm, DNA algorithm, Enhanced Map-Reduce, Cloud Computing.*

1. Introduction

Classification has been identified as the backbone functionality under data mining field to handle massive amount of data produced on a daily basis [4]. Today's world widely accepted domains like speech recognition, handwriting recognition, document classification etc. have accepted classification as an efficient approach to handle the data. The massive data generated on a daily basis comes with many challenges dealing with handling the data, retrieving the quality of data, storage, and availability and so on. In this paper, the main focus is regarding the medical dataset present in the form of text which consists of large amount of sensitive data regarding all the medical records, the financial data, insurance data etc. which requires high security and maintenance [7]. The data present over the cloud is at a high risk rate of getting manipulated by the third parties which can cause a severe impact on a human's life [11]. For example, different organizations have to maintain and format the health care data to match the patient records which can raise some uncertain conditions such as data loss or data breach over the public cloud[8][11]. This paper has looked majorly over the following aspects:

1. Extracting data and data security over public cloud. Taking into consideration the first aspect that is segregating out the relevant information out of the raw dataset is a challenging task which can be approached in an other way round using unsupervised form of learning K-means clustering [1][4]. Though real time analysis of data is done under unsupervised learning, KNN classification plays an efficient role that focuses on off-line analysis resulting into more accurate and reliable outcomes and thus also depicts the advantage of having the training classes known before hand [2][6].Efficient management of data can result into cost efficiency and optimal performance.

2. Looking into the other aspect for data security, when the data has to be outsourced to the public cloud, here raises a privacy concern to store and maintain the data over the cloud. This project aims to solve the above problem by making use of Enhanced Map-Reduce in addition to Classification along with the unique Hash Code generation which makes the optimal security of the

May-June 2020
ISSN: 0193-4120 Page No. 4303-4308

data stored over the cloud [5][9][10].

3. The last aspect talks about the authentication of data focusing on encryption and decryption which can be achieved through DNA encryption as an optimal method for storing and retrieving the data [3].

Further the paper includes the following as below. In section I, the review of the related work has been given. The section II describes about the objectives of the proposed model. In section III the proposed system is demonstrated. The section IV includes the modules identified in the proposed system. Further the working behind the proposed model is shown in the section V. The next section VI displays the results and the section VII concludes the paper.

2. Related Work

The process of handling the big data and providing security for the data has always been a major concern in the modern world. In the previous years, a lot of research and studies are done in order to easily access and avail the data. An addition to the study about protecting the data over the cloud has also been a concerned field of study.

- **"Practical Privacy-Preserving Map Reduce Based K-means Clustering over Large-scale Dataset" [1]**This paper was presented by **Jiawei Yuan, Member, IEEE, Yifan Tian, Student Member, IEEE** which proposed a model where data undergone through K-means clustering can be outsourced to the cloud platforms. The scheme though involves a major disadvantage of low transmission speed and less security over the cloud as it makes use of the threat model.

- **"A Comparative Study on Clustering and Classification Algorithms"[2]**presented by **Jyotismita Go swami Assistant Professor, Department of Computer Science ,Arunachal University of Studies Namsai, Arunachal Pradesh, India** has laid down important aspects of machine learning to study the nature of data and classify using different methods of learning.

- **"A Top-k query algorithm for big data based on Map Reduce" [3]** proposed by **X. Lin** .This paper conveys the research on a highly efficient technique as Top k-query in the context to achieve optimal measures of data portioning and data reduce.

3. Objectives

The main objective of the proposed model is to provide an efficient way for textual data classification and further to the security of the data to be uploaded over the public cloud. The objectives focus on the following:

- Textual data.
- Text Pre-processing.
- Text classification-KNN classification.
- Overcoming the drawbacks of K-means clustering.
- Making use of the Enhanced Map Reduce including two aspects- Divide & Conquer and Hash code generation.
- Privacy concern over public cloud environment with the help of DNA algorithm.

4. Proposed System

We are proposing a system with KNN classification over Large-scale Dataset using Enhanced Map Reduce technique and DNA Encryption algorithm.

First stage includes initializing trained data set for every different cluster which is related to medical Information. After, the clustering algorithm divides the file into number of blocks and for every block hash code is generated for the security purpose. Before storing into cloud platform, classification algorithm classifies that file belong to which cluster category by matching it with the training clusters already stored over the public cloud.

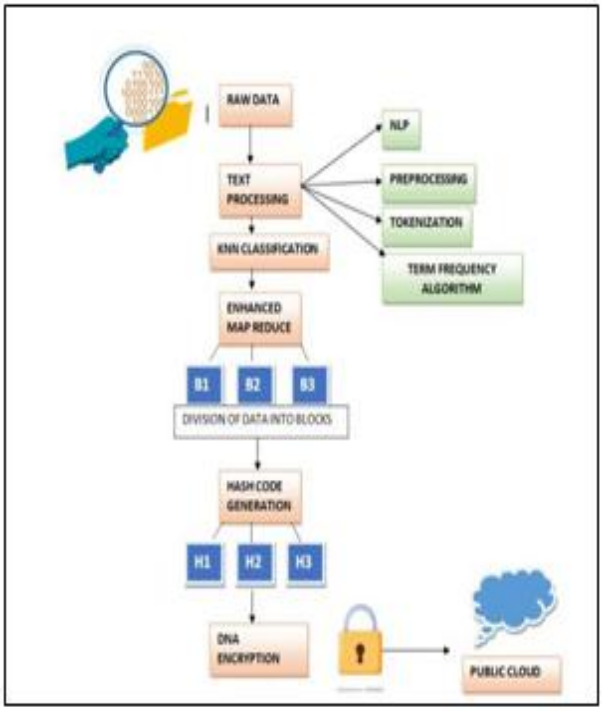

Figure 1: System Architecture

5. Modules

The proposed system is segregated into various modules where every module holds mechanism to overcome the challenges to handle the big data. The modules identified are given as below:

May-June 2020
ISSN: 0193-4120 Page No. 4303-4308

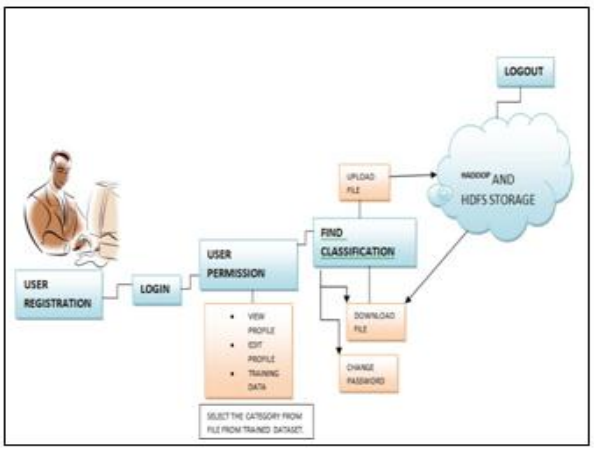

Figure 2: Modules Identified

➢ Textual Data

Select the cluster category from the dataset and store it in database based on selected clusterid.

➢ Upload File Module

Select the file from local system and click to upload option.

➢ Pre-processing

1. In the selected file have to remove unnecessary words.

2. Comparing file content with trained dataset. If it is matching with trained dataset then increasing the count of category code (Cluster id). Which category code having max count that file is belongs to that category(Cluster).

3. User already selected file has to store in that cluster.

4. Select file from cluster table then selected files will get divided into small blocks (500 bytes each block) and each block content will get encryption by using DNA Algorithm (EncryptionKey) [3].

eg: packet size=500;

File Size=3000

=3000/500

=6 blocks

5. Generate hash tag for all block.

- Compare generated hash block with existing hash tag from database if hash tag matched in that case we will not upload that block into hadoop, we will increase number of instance of that block in database able [9].
- If hash tag not matched in that case we will add that block hash details in database and upload that block in hadoop.
- LBA - Logical Block Addressing technique is used to identify what are the blocks are present in a file.

➢ Download File Module

- Select the file in download list.
- Get the LBA based on file id.
- Each encrypted blocks has to get decrypt by using DNA algorithm.
- The file to be downloaded can be selected by the user.
- Using LBA has to find block numbers which are in selected file.
- Whether all the blocks required for the file is available, if all the blocks are available in Hadoop storage space and download blocks, while downloading itself all the encrypted blocks will get decrypt by using DNA algorithm (Decrypt Key) then merge the blocks and give it to the user.

6. Methodology

In the proposed system, we have a client end and a server end where the client end refers to the raw data provider and the server end refers to the public cloud. The outsourcing of the data is done from the client end to the server end i.e. the public cloud. The different stages involved in outsourcing the data are as follows:

Client-end:

Stage 1: Input as raw data

The raw data which is taken as an input consists of textual medical data consisting of medical terminologies. *For example: A text document consisting of the report of a cardiac patient.*

Stage 2: Text Processing

Text processing involves various stages starting from NLP, Pre-processing, Tokenization and Term Frequency Algorithm which helps in understanding of spoken or written terminologies through the medium of a computer[7]. The text pre-processing makes the data into a more digestible form and the long textual data is further divided into chunks or small pieces of data with the help of tokenization. Further the term frequency algorithm checks frequency of data occurrences.

Stage 3: K-Nearest Neighbor Classification KNN Algorithm:

- KNN algorithm is a supervised form of machine learning also known as lazy learning algorithm which helps to solve classification problems [2].
- First step is to determine the no of nearest neighbors i.e. k.
- Calculate the distance between the new data and the training data using Euclidean's formula.

$$D(p,q)=\sqrt{(q1-p1)^2+(q2-p2)^2}$$

- New data is (Y), thus the new data Y belongs to the k means nearest training class.

Stage 4: Enhanced Map Reduce Map-Reduce Algorithm:

Map Reduce-a functional programming model proceeds by dividing the input task further into smaller and manageable sub-tasks which should be independently executable. The tasks can be worked out in-parallel to each other [5].

May-June 2020
ISSN: 0193-4120 Page No. 4303-4308

- Map Function-Splitting & Mapping.
- Shuffle Function-Merging & Sorting.
- Reduce Function-Reduce step (final step).

Hash-code generation:

Hashing is the process of converting a string of characters into a short fixed-length block which is the other form of the original string, which is known as hash code. Hashing is mainly used for indexing [10].

Here hashing is used for comparing the blocks which are generated and the blocks which are already present in the cloud. By doing so we can make out whether the block is already present or not, if it's present indexing will take place. There are different types of hashing algorithm, but here we use MD5 algorithm. The MD5 function is a cryptographic algorithm that converts variable length input into a message digest or hash code which is of 128 bits long [9][10].

MD5 Algorithm:

1. Append Padding bits-The input which is given is padded with extra bits so that it is congruent to 448. Modulo512.Which means it is extended to 64 bits shy of being 512 bits long.

2. Append the length-A 64 bit representation of b is appended to result of previous step. The final result will be of the exact multiple of512bits.

3. Assign the Initial value to MD Buffer.

Here 4-Word buffer is used to compute the message digest i.e.(A,B,C,D)each is of 32 bit register.

A=01 23 45 67, B=89 ab cd ef,

C=fe dc ba 98, D=76 54 32 10

4. Processing of input into 16-wordblocks.

5. Final output-The message digest produced as output is A, B, C, and D.

Server end:

Stage 5: DNA Algorithm

The DNA implementation involves two phases [3]:

1. Encryption of Secret Data
2. Extracting the Original Data.

Phase 1:

1. The encryption process begins with conversion of binary data to DNA sequences.

A=00, T=01, C=10, and G=11.

2. The next step is the complementary rule in which a unique equivalent pair is assigned to every nucleotides base pair.

Complementary rule: ((AC) (CG) (GT) (TA))

Example: DNA strand: AATGCT
After applying Complementary rule: CCATGA.

3. Numerical data which is represented using DNA sequence.

DNA Reference Sequence:

[TG_{00},TA_{01},AT_{02},GC_{03},CT_{04},GA_{05},CA_{06},AC_{07},AA_{08}, GT_{09},CG_{10},AG_{11},CC_{12},TT_{13},TC_{14},GG_{15}]

At the client end, the original data M has to uploaded via a network to cloud server 1 which undergoes three sub-phases with the outcome as final form of M which is M´´´ and that is further uploaded to cloud server 2.

For example:
M=10000010
Applying base pair rule:
(A= 00, T= 01, C= 10, G= 11)□M'=CAAC
Applying complementary rule:
((AC) (CG) (GT) (TA))□M"=GCCG
After applying numeric indexing:0310

Phase 2:

The decryption phase takes place at cloud server 2 where the encrypted data is present. To get back the original data it has to be decrypted using the following three sub-phases:

1. Conversion of numeric data i.e. encrypted data into DNA Sequences.DNA reference sequencing is done according to the index read from the file.

2. The next step is the complementary rule in which a unique equivalent pair is assigned to every nucleotides base pair.

3. Conversion of DNA sequences to binary data.

For example: M'''=0310
Referencing DNA numeric indexing: M"=GCCG
Applying complementary rule:
((AC) (CG) (GT) (TA))M'=CAAC
And the original data:M=10000010.

DNA Implementation:

The DNA implementation can be done in two different ways as follows:

1. Symmetric cryptography
2. Asymmetric cryptography

The DNA Encryption is as follows:

1. The input data and the key is converted into ASCII bits.

2. The ASCII bits which are generated are not even, so to make it even zero padding will be done.

3. Then the XOR operation is done on the input and key.

4. The output of the XOR operation is represented using DNA bases format .This format is now known as enciphered text. The representation is shown in Phase 1

May-June 2020
ISSN: 0193-4120 Page No. 4303-4308

and the encryption model is shown in figure 3.

5. Likewise the decryption as shown in figure 4 also takes place but in a reverse order.

Figure 3: Encryption model

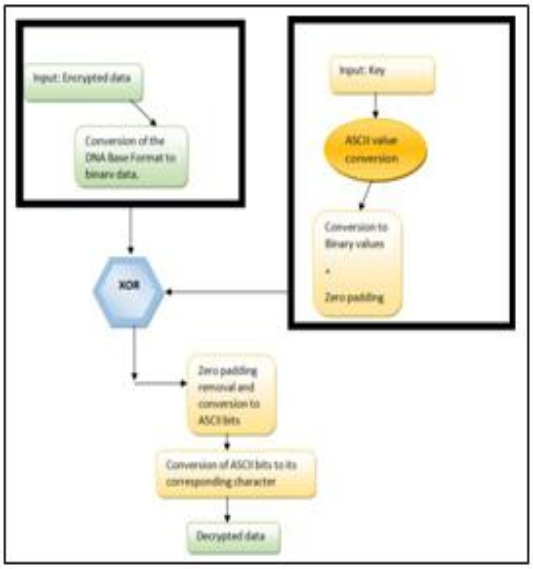

Figure 4: Decryption model

7. Results and Discussions

In the proposed working model for the textual data classification, the enhanced map reduce has played an important role bringing about the ease of handling data and eliminating the unwanted matter. The DNA algorithm designed to overcome the privacy concerns over the public cloud has helped sportingly the proposed system with high efficiency and effective storage over the cloud platform.

Figure 5: Home Page

- The project has a user interface where collection of different medical datasets is available. A user can easily upload raw text data into the cloud by undergoing the text processing methodologies.

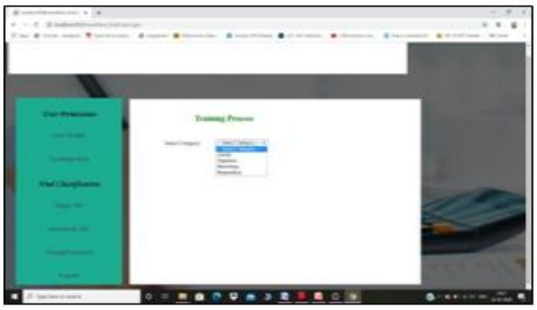

Figure 6: Uploading and Downloading File

The raw data gets compared to the training class data available in the cloud with the help of the hash code algorithm. The hash code index is referenced if present otherwise a new cluster category is formed. The user can download the relevant data as required.

- The encryption and decryption takes place with the help of the DNA algorithm under this project.
- Thus it helps an efficient and secure way to categorize data as well as storage and retrieval of data.

8. Conclusion and Future Scope

In this paper, Enhanced Classification and DNA security on Large Scale Dataset is proposed on the cloud computing. Here DNA implementation is used for better encryption process overcoming the drawbacks of the existing work which had the concept of AES and DES algorithms. The AES and DES algorithms being widely accepted in the encryption process, also hold back the limitation of getting more prone to attacks like some modern technologies has been developed to break the AES and DES algorithm. Hence DNA cryptography overcomes the above problem by developing an unbreakable algorithm.

This project achieves classification speed and accuracy is more than the existing work. KNN classification also assures the time complexity factors and easy classification of data. Privacy concern solved with

May-June 2020
ISSN: 0193-4120 Page No. 4303-4308

the help of the hashing algorithm which produces better performance measures and the user-cloud relationship user friendly. Here thorough analysis is done on the security and efficiency. In reference to future scope still many upcoming technologies can help in better security over the public cloud thus maintaining the integrity of the data. The proposed model can be further designed to access multiple file over the cloud with enhanced security.

References

[1] Jiawei Yuan and Yifan Tian," Practical Privacy-Preserving Map Reduce Based K-means Clustering over Large-scale Dataset" IEEE Transactions on Cloud Computing (Volume: 7, Issue: 2 , April-June 12019).

[2] A. Moldagulova and R. B. Sulaiman, "Using KNN algorithm for classification of textual documents," 2017 8th International Conference on Information Technology (ICIT), Amman, 2017, pp.665-671.

[3] Vinay kumar Pant and Ashutosh Kumar "DNA Cryptography An New Approach to Secure Cloud Data" International Journal of Scientific & Engineering Research, Volume 7, Issue 6, June-2016 ISSN2229-5518.

[4] Jyotismita Go swami "A Comparative Study on Clustering and Classification Algorithms" International Journal of Scientific Engineering and Applied Science (IJSEAS) - Volume-1, Issue-3, June 2015 ISSN: 2395-3470.

[5] X. Lin, "A Top-k query algorithm for big data based on Map Reduce," *2015 6th IEEE International Conference on Software Engineering and Service Science (ICSESS)*, Beijing, 2015, pp.982-985.

[6] G. Aizhang and Y. Tao, "Based on Rough Sets and the Associated Analysis of KNN Text Classification Research," 2015 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES), Guiyang, 2015, pp.485-488.

[7] F. Amato et al., "Challenge: Processing web texts for classifying job offers," Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), Anaheim, CA, 2015, pp.460-463.

[8] A. Hendre and K. P. Joshi, "A Semantic Approach to Cloud Security and Compliance," 2015 IEEE 8th International Conference on Cloud Computing, New York, NY, 2015, pp.1081-1084.

[9] Piyush Gupta, Sandeep Kumar "A Comparative Analysis of SHA and MD5 Algorithm." International Journal of Information Technology and Computer Science 5(3):4492 - 4495 · June 2014 *with* 3,863Reads.

[10] R. Roshdy, M. Fouad and M. Aboul-Dahab "Design And Implementation A New Security Hash Algorithm Based On MD5 And SHA-256" International Journal of Engineering Sciences & Emerging Technologies, August 2013. ISSN:2231- 6604 Volume 6, Issue 1, pp: 29-36 ©IJESET.

[11] D. R. Bharadwaj, A. Bhattacharya and M. Chakkaravarthy, "Cloud Threat Defense – A Threat Protection and Security Compliance Solution,"2018 IEEE International Conference on Cloud Computing in Emerging Markets (CCEM), Bangalore, India, 2018, pp.95-99.