International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 04 | Apr-2018 www.irjet.net p-ISSN: 2395-0072

A DATA MINING WITH BIG DATA DISEASE PREDICTION

Yamini.S[1], Rama prabha.K.P[2]

[1]*Student, MS [software engineering]*
[2]*Professor, School of Information Technology & Engineering, VIT University, Vellore, Tamilnadu, India*

Abstract - *Global Healthcare industry is generally having large amount of data in Knowledge rich but unfortunately not all the data are classified, analyzed and mined which is required for discovering hidden pattern which is used for find out disease and providing solution using Data mining Machine Learning Techniques for classification ,managing privacy using advance encryption standard(AES) and Advance Big data Techniques sqoop for data migration and MapReduce used for Analysis and clustering this process refers and develops a new methodology using Hadoop, based upon this processed data using d discover knowledge in database and for medical research, particularly in disease prediction. This paper aims at classifying, analyzing and processing the various medical data from the global using big data and data mining techniques. Also targets utilization of large volumes of medical data while combining multimodal data from disparate sources.*

Key Words: **Data mining, Machine Learning, AES, Big data, Sqoop, MapReduce, Cluster.**

1.INTRODUCTION

In this privacy patient clinical decision support system which allows service provider to diagnose patient's disease without leaking any patient's medical data. In this system, the past patient's historical medical data can be used by service provider to train the Cluster based analysis and prediction. It creates unique key for each user in the system, and also it provides encrypted data for providing security. Then, service provider can use the trained classified group to diagnose patient's diseases according to his symptoms in a privacy and preserving way. Finally, patients can retrieve the diagnosed results according to his own preference privately without compromising the service provider's privacy. The dataset are collected from this process that can be imported to the Hadoop for data migration through Sqoop. Finally the data can be applied to MapReduce process for clustering data.

1.1Predictable attribute

- Disease
- Diagnosis
- Treatment

1.2 Input attribute

- Admin

(updating doctor and provider details)

- Doctor

(Providing solution for patient queries)

- Provider

(Uploading Medicine related information for process)

- User/Patient

(Symptoms Selection / Sending Messages for knowing solution)

- Symptoms

(<Symptom 1><Symptom 2><Symptom 3>initial stage to last stage of symptoms can be selected)

2. DATA MINING MACHINE LEARNING ALGORITHM

The Naive Bayesian Classifier technique is processed when the dimensional of the inputs is high. Its simplicity, also it can highly sophisticated method for classification. It model recognizes the characteristics of patients with disease. It displays the probability of symptoms for the predictable state. It is the basis for many machine learning and data mining method. This is used to create models with predictive capabilities also provides new ways of exploring and understanding data.

Probability of the conclusion say C=observation, E= relationship exists between C & E.

This probability is = P (C |E) where P(C|E) = P(E|C) P(C) divided by P(E)

Working progress of Bayes rule:

[1] D= Training set of tuples and Ca & Cp =Associated class labels. Record of the system each has been represented like n-dimensional AV(attribute vector),Y=(y1, y2…,yn-1, yn), n measurements made on the tuple from n attributes. A1 to An.

[2] Suppose that there are m number of classes for prediction, C1, C2… Cm. Given a record, Y, the classifier will predict that Y belongs to the class having the highest posterior probability, conditioned on Y. That is, the naive Bayesian classifier predicts that tuple x belongs to the class Ci if and only if P (Ci|Y)>P (Cj|Y) for 1≤ j≤m and j ≠ i.Thus we maximize P(Ci|Y). The class Ci for which P (Ci|Y) is maximized is called the maximum posteriori hypothesis. By Bayes' theorem P(Ci|Y) = P(y|Ci)P(Ci) divided by P(x).

[3] As P(Y) is constant for all classes, only P (Y|Ci)* P (Ci) need be maximized.in this process class prior probabilities are not knowing, then it is assumed that the classes are equally likely, that is, P (C1) =P (C2) =…P(Cm -1)=P (Cm) and

we would therefore maximize P (Y|Ci). Otherwise, we maximize P (Y|Ci) P (Ci). Note that the class prior probabilities may be estimated by P (Ci) =|Ci, D|/|D|, where |Ci, D| is the number of training tuples of class Ci in D.

[4] Data sets with multiple number of attributes would be expensive to compute P(Y|Ci). To reducing computation in evaluating P(Y|Ci), the naive assumption of class conditional independence is made. This presumes that the values of the attributes are conditionally independent of one another, given the class label of the tuple (i.e., that there are no dependence relationships among the attributes). Thus, P(Y|Ci)=IIm$^{K=1}$ P (yk|Ci) =P (y1|Ci) * P (y2|Ci) *… P (ym|Ci).

We can easily estimate the probabilities P (y1|Ci), P (y2|C i)… P (ym|Ci) from the database training tuples. Recall that here yk refers to the value of attribute Ak for tuple Y. For each attribute, we will see that whether the attribute is categorical or continuous-valued. For instance, to compute P (Y|Ci), we consider the following: If Ak is categorical, then P (Yk|Ci) is the number of tuples of class Ci in D having the value yk for Ak, divided by |Ci, D|, the number of tuples of class Ci in D.

[5] In order to predict the class label of Y, P(Y|Ci)P(Ci) is evaluated for each class Ci. The classifier predicts that the class label of tuple Y is the class Ci if and only if P(Y|Ci)P(Ci)>P(Y|Cj)P(Cj) for $1 \le j \le m, j \ne i$,In this process predicted class label is the class Ci for which P (Y|Ci) P (Ci) is the maximum.

In this system selecting symptoms process phase this technique can be applied and finally the disease, diagnosis and treatment predicted.

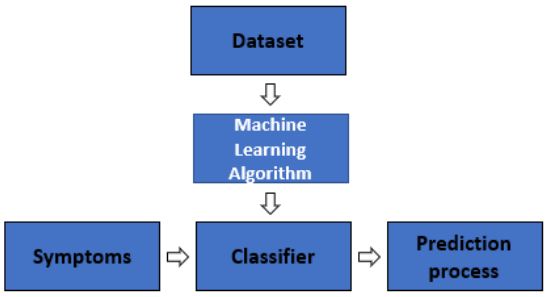

Fig -1: Classification

3.ADVANCE ENCRYPTION STANDARD

The Symptoms data in the system, processor of the data begins by index and it encrypts with a symmetric encryption scheme under a unique key (UK) process. It then encrypts the index using a searchable encryption scheme and encrypts the UK with an attribute-based encryption scheme. At the final it encodes the encrypted data and index in a way that the data verifier can verifies their integrity (proof of storage), protect classified information.

Working progress of AES

[1] The process basically selection of symmetric key progressing algorithm.
[2] Transparent analysis of the designs processed. Advance Encryption Standard blocks cipher.
[3] It capable of handling 128-bit blocks, using keys sized at 128, 192, and 256 bits.

Advantages

- Security
- Cost
- Simplicity of implementation

4.BIG DATA

The Big data is categories as 3Vs (Volume, Variety, Velocity) and also it has Veracity. Most of the data is unstructured, quasi structured or semi structured and it is heterogeneous in nature. The volume and the heterogeneity of data with the speed it is generated, makes it difficult for the present computing infrastructure to handle Big Data. Traditional data management, warehousing and study systems fall short of tools to analyze this data. Due to its precise nature of Big Data, it is stored in distributed file system architectures.
Working Progress of Big data

4.1 Sqoop

Sqoop is a big data tool that process the capability to extract data from non-Hadoop data stores or traditional database system, transform the data into a form usable by Hadoop, and then load the data into hadoop distributed file system. This process is called Extract, Transform, and Load(ETL).

4.2 Hadoop

Hadoop is set of tools that supports running of applications on Big Data. Hadoop addresses the challenges in the Big Data. supports distributed applications, and allows the distributed processing of bulk data sets through the commodity servers. . Hadoop implements map-reduce using Hadoop Distributed File System.

4.3 MapReduce

MapReduce is famous for large amount of data processing and analysis of voluminous datasets in clusters of machines.

- **Map Phase**

Initially split the data into key value pair and fed into mapper which in turn process each key value pair and generate intermediate output.

- **Reduce Phase**

The Intermediate key value pair first collected sorted and grouped by key and generates values associated with each key. The receiver produces final output based on some calculation and stores it in an output file.

The Map and Reduce task runs as sequential process in cluster; the output of the Map part becomes the input for the Reduce part. From this technique can be applied into the data and the disease process can be applied this system basically process single node cluster.

5.SYSTEM ARCHITECTURE

In this system admin update the physician and trust authorizes register. The trust authorizes can login and input the symptoms, it could be classified and disease predicted and treatment suggested. From this process while dataset collected data can be imported traditional system into big data system through sqoop. MapReduce process will be applied to the dataset for clustering process.

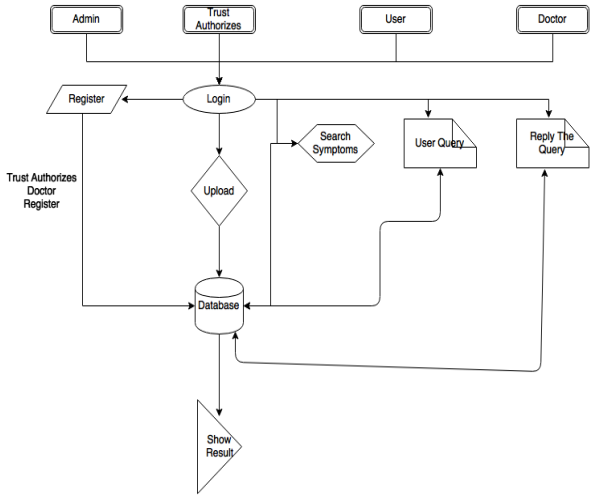

Fig -2: Disease Prediction

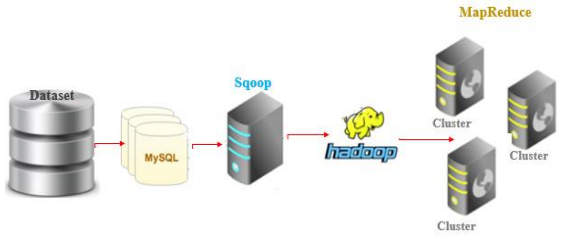

Fig -3: Data Clustering

6.SYSTEM RESULTS

A popular data-mining technique that is used to classify a dependent variable based on measurements using naive bayes classification technique disease can be predicted the classified data can be imported using Sqoop process from traditional database into Hadoop. Hadoop implements MapReduce**,** The symptoms or query can be processed based upon that disease can be predicted, using classification. And dataset will be imported into Hadoop then the result be in the form of clusters.

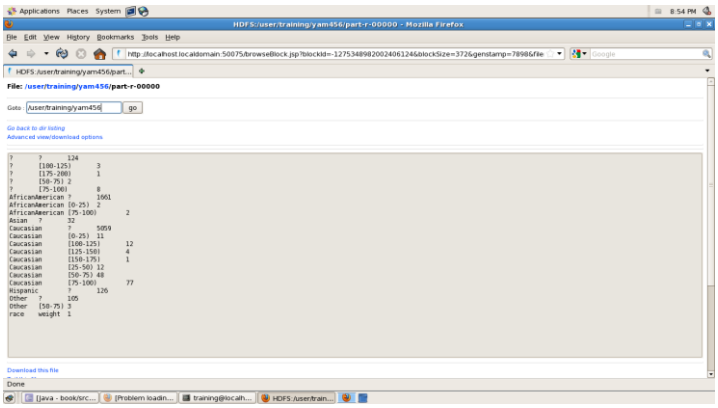

Fig -4: Cluster

7.CONCLUSION

Proposed a system using Analysis and Clustering based prediction. System can use big medical dataset stored and then apply the cluster for disease diagnosis without compromising the privacy of Disease Prediction. In addition, the patient can securely retrieve the top-k diagnosis results according to their own preference in our system. Since all the data are processed in the encrypted form, our system can achieve patient-centric diagnose result retrieval in privacy preserving way. In this system diagnose patient's disease without leaking any patient's medical data. The past patient's historical medical data can be used by service provider to train the classification and Cluster based analysis prediction. In this system follows single node cluster technique. For the future work, I will exploit this system with advanced Data mining and Big data techniques in Multi-Node Cluster way which could process Data replication.

REFERENCES

[1] H. Monkaresi, R. A. Calvo, and H. Yan, "A machine learning approach to improve contactless heart rate monitoring using a webcam," IEEEJ. Biomed. Health Informat., vol. 18, no. 4, pp. 1153–1160, Jul. 2014.

[2] C. Schurink, P. Lucas, I. Hoepelman, and M. Bonten, "Computer-assisted decision support for the diagnosis and treatment of infectious iseases in intensive care units," Lancet Infectious Dis., vol. 5, no. 5, pp. 305–312, 2005.

[3] I. Kononenko, "Machine learning for medical diagnosis: History, state of the art and perspective," Artif. Intell. Med., vol. 23, no. 1, pp. 89–109, 2001.

[4] N. Lavraˇc, I. Kononenko, E. Keravnou, M. Kukar, and B. Zupan, "Intelligent data analysis for medical diagnosis: Using machine learning and temporal abstraction," Artif. Intell. Commun., vol. 11, no. 3, pp. 191–218, 1998.

[5] Y. Elmehdwi, B. K. Samanthula, andW. Jiang, "Secure k-nearest neighbor query over encrypted data in outsourced environments," in Proc. IEEE 30th Int. Conf. Data Eng., pp. 664–675, 2014.

[6] P. Paillier, "Public-key cryptosystems based on composite degree residuosity classes," in Proc. Adv. Cryptol. Int. Conf. Theory Appl. Cryptograp. Techn., Prague, Czech Republic, May 2–6, 1999, pp. 223–238.

[7] Y. Elmehdwi, B. K. Samanthula, and W. Jiang, "k-nearest neighbor classification over semantically secure encrypted relational data," IEEE Trans. Knowledge Data Eng., (2015)..

[8] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," Int. J. Med. Informat., vol. 77, 2008.

[9] Judith Hurwitz, Alan Nugent, Dr. Fern Halper, and Marcia Kaufman, "Big Data for Dummies", John Wiley & Sons, Inc., 2013

[10] Prajapati, Vignesh, "Big Data Analytics with R and Hadoop", Packt publishing Nov 2013

[11] Jyoti Soni, Ujma Ansari, Dipesh Sharma, Sunita Soni," Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction", International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[12] Data mining concepts and techniques, second edition, Han Kamber.

BIOGRAPHIES

Yamini.S
Student
MS[Software Engineering]
VIT University,Vellore

Rama prabha K.P
Assistant Professor
School of Information Technology and Engineering
VIT University, Vellore