Clinical Decision Support System for Privacy Preserving using Information Retrieval

Student Ms. Pradnya Kul[1], Dr. V. S. Bidve[2]

Department of Information Technology

Smt.Kashibai Navale College of Engineering, Savitribai Phule Pune University, Pune,India

Abstract *—In this paper, the proposed system is designed which predicts the accurate disease and prevention. Information retrieval techniques are used such as data cleaning, data smoothing, data clustering to get the data required for prediction. To get accurate prediction and prevention of disease K-means algorithm is used in the system. It basically partitions the data into cluster and then finds the result. In addition performance criteria via extensive simulation also demonstrate that the system can effectively calculate patient's disease risk with high accuracy in privacy preserving way. All this data is stored in cloud with encryption technique. So result for privacy preserving is more accurate*

Keywords — *Clinical decision Support System, Patient centric, Naive Bayesian classifier, K-means clustering, AWS S3*

I. INTRODUCTION

Clinical Decision System (CDS) is the system used to diagnosis the patient in different way. It mainly check the patient symptoms, test result and historical data. To predict the disease or treatment there are many more algorithms used based on medical phenomenon. To speed up the diagnosis time and improve the diagnosis accuracy, a new system in healthcare industry should be workable to provide a much cheaper and faster way for diagnosis.[7] Naïve Bayesian Classifier, one of the popular machine learning tools, has been widely used recently to predict various diseases in Clinical Decision Support System (CDSS). It performs well in multi class prediction.

The past patient's historical data are stored in cloud and can be used to train the naive Bayesian classifier without leaking any patient's historical /medical data and then the trained classifier can be applied to compute the disease risk for new coming patients and also, allow these patients to retrieve the top-k disease names according to their own preferences.[7] To overcome the machine learning strategy k-means clustering algorithm is used. It basically clusters the disease and predict, what treatment should be do next.

The proposed system discussed in this paper is to improve the performance of the clinical decision support system and to improve the diagnosis time and accuracy.

A. Machine Learning

Machine itself learns the strategy of given input and output also it shows the output for new input based on given data set.

There are mainly two types of machine learning.

Supervised learning: Machine is presented with example of input and their desired output. Using this method machine learns the general rule.

Unsupervised learning: Machine doesn't know the input and its desired output. Machine itself learns the algorithm and produce the output.

B. Naive Bayesian classifier

Bayesian classifier could represent the probabilistic relationships between diseases and symptoms. Naive Bayesian classifier is a classifier which has been proved to be effective in many practical applications, including text classification, medical diagnosis, and systems performance management.

Naive Bayes theorem provides a way of calculating prediction of risk disease, P(c|x), from P(c), P(x), and P(x|c).

Naive Bayes classifier assume that the effect of the value of a predictor (x) on a given class (c) is independent of the values of other predictors.

$$P(Ci \mid X) = P(X \mid Ci)P(Ci)/P(X)$$

- P(c|x) is the probability of class (target) given predictor (attribute).
- P(c) is the probability of class.
- P(x|c) is the likelihood which is the probability of predictor given class.
- P(x) is the prior probability of predictor.

C. K-means clustering algorithm

Cluster analysis is the phenomenon of a set of observations into clusters so that observations within the same cluster are similar according to some desire criteria, while observations from different clusters are dissimilar.

k-means clustering algorithm is mainly used for partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the disease

Step 1. Randomly select "c" cluster centers.

Step 2. Calculate the distance between each data point and cluster centers.
Step 3. Assign the disease (data) to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
Step 4. Recalculate the new cluster center using:

$$ci = (1/mi)\ x\ ci\ x=1$$

=1

Where...ci" represents the number of data points in ith cluster.
Step 5. Recalculate the distance between each (data) disease and new obtained cluster centres.
Step 6. If no data point was reassigned then stop, otherwise repeat from step 3.

D. Cryptography Algorithm

1. Paillier Homomorphic Encryption Algorithm
Input: A plain text message
Output: An encrypted message for multiple participants
Step 1: Initialize the variables p1,p2,…..pn andq1,q2,…qnsuch that they are large primes.
Step 2: Generate the public key for each participantn1,n2,….nk such that n=p*q
Step 3: Choose the semi random variable "g"
Step 4: Calculate the encrypted value for each participant by the formula

$$Ci = gxi \cdot rini \bmod ni2$$

2. Paillier Homomorphic Decryption Algorithm
Input: An encrypted message divides among many participants ci.
Output: the plain text message among the participants.
Step 1: Calculate the value of λ by the LCM of the pi and qi.
Step 2: Evaluate the value of "u"by applying the formula

$$ui= g\ \lambda(ni) \cdot rini \bmod ni2$$

Step 3: Calculate the inverse of the L(ui) which will give you the plain text value.

E. AWS S3 Cloud

In this strategy storing the historical, medical patient's data cloud is used. Amazon Web Service S3 has simple web services interface that we use to store and retrieve any amount of data, at any time, from anywhere on the web. AWS provides secured way to stored data on cloud with retrieve and manage the permission of resource. All this storage process is followed by the authentication and access control method. In this process access control gives the authority that who can access objects with type of access (READ,WRITE).

Following are some basic element in AWS:
1.Bucket: A bucket is a container for objects stored in AWS S3. Every object is contained in a bucket. It organize the AWS S3 namespace at the highest level, they identify the account responsible for storage and data transfer charges, they play a role in access control, and they serve as the unit of aggregation for usage reporting.
2.Objects: Objects are the fundamental entities stored in Amazon S3. Objects consist of object data and metadata. The data portion is opaque to AWS S3.
3.Key: A key is the unique identifier for an object within a bucket. Every object in a bucket has exactly one key. Because the combination of a bucket, key, and version ID uniquely identify each object.
4.Region: We can choose the geographical region where AWS S3 will store the buckets created. User can choose a region to optimize latency, minimize costs, or address regulatory requirements.

Following are the common operation in AWS:
1. Create a Bucket – Create and name your own bucket in which to store your objects.
2. Write an Object – Store data by creating or overwriting an object. When you write an object, you specify a unique key in the namespace of your bucket. This is also a good time to specify any access control you want on the object.
3. Read an Object – Read data back. You can download the data via HTTP
4. Deleting an Object – Delete some of your data.
5. Listing Keys – List the keys contained in one of your buckets. You can filter the key list based on a prefix.

Features of AWS S3:
Object store model for storing, listing, and retrieving data.
Support for objects up to 5TB, with many bytes of data allowed in a single bucket.
Readable and writeable from Apache Hadoop, Apache Spark, and related applications.
Readable and writeable from other applications.

F. System architecture:

Clinician enters the historical data of respective patient and data provider encrypts that data. When there is requirement of historical data hospital decrypt that data and make the prediction of health issue using naive Bayesian classifier. Clinician enters the data about current status of patient and all this data is encrypted. After that clustering algorithm works for clustering the prediction of disease. Using result of K-means doctors recommend the prescription and further treatment.

II. LITERATURE SURVEY

Many techniques have been widely used in clinical decision support systems for prediction and diagnosis of various diseases with better accuracy. These techniques have been very effective in developing clinical support systems because they are able to detect hidden patterns and relationships in medical data. Classification is a data mining techniques that assigns items in a collection to target categories or classes. The aim of classification is to

predict the target class for each case in the data accurately.

There are various approaches used to predict the disease. Various data retrieval techniques have been used for disease prediction demand. The work done for prediction of disease is discussed below:

A. *Predictive data mining technique*

A lot of work is done in the field of Clinical decision support system to enhance the diagnosis of disease. In the process of prediction of disease there are many more symptoms and test result are important. Patient's symptoms and historical health are main constraints at the time of prediction.

R. Bellazzi and B. Zupan: Theses authors have introduced "Predictive data mining in clinical medicine Current issues and guidelines". This review is mainly to discuss extent of strategy and role of the research area i.e. clinical data lof predictive data mining. It also have the framework to cope with the problems of constructing, assessing and exploiting data mining models in clinical medicine.[1]

B. *Gaussian Kernel based classification*

Y. Rahulamathavan, S. Veluru, R. Phan, J. Chambers, and M. Rajarajan: Authors introduced "Privacy-Preserving Clinical Decision Support System using Gaussian Kernel based Classification". This paper has proposed a privacy-preserving decision support system using a Gaussian kernel based support vector machine approach. Since the proposed algorithm is a potential application of emerging outsourcing techniques such as cloud computing technology, rich clinical data sets available in remote locations could be used by any clinicians via the Internet without compromising privacy, thereby enhancing the decision making ability of healthcare professionals.[2]

C. *Privacy Preservation in data storage*

Y. Tong, J. Sun, S. S. M. Chow, and P. Li: introduced "Cloud-Assisted Mobile- Access of Health Data with Privacy and Audit ability". The author has proposed a system to build privacy into mobile health care systems with the help of the private cloud. We provided a solution for privacy-preserving data storage by integrating a PRF based key management for unlike ability, a search and access pattern hiding scheme based on redundancy, and a secure indexing method for privacy-preserving keyword search. This paper have investigated techniques that provide access control and audit ability of the authorized parties to prevent misbehaviour, by combining ABE-controlled threshold signing with role-based encryption.[3]

D. *Big Data Technique*

Rongxing Lu, Hui Zhu, Ximeng Liu, Joseph K. Liu, and Jun Shao introduced "Toward Efficient and Privacy-Preserving Computing in Big Data Era". Paper has investigated the privacy challenges in the big data phenomenon by first identifying big data privacy requirements. They have also introduced an efficient and privacy-preserving similarity computing protocol in response to the efficiency and privacy requirements of data mining in the big data era.[4]

E. *Machine Learning approach*

H. Monkaresi, R. A. Calvo, and H. Yan authors introduced "A Machine Learning Approach to Improve Contactless Heart Rate Monitoring Using a Webcam". They evaluated a method for remote Heart Rate measuring in three applications: a controlled laboratory task, a naturalistic HCI, and an indoor cycling exercise are the three method for measuring jeart rate. This study evaluated Poh et al.'s method and showed the feasibility of their methodology to measure HR at rest.[5]

F. *Privacy-Preserving using Computation Method*

Erman Ayday, Jean Louis Raisaro, Paul J. McLaren, Jacques Fellay, Jean-pierre Hubaux these author has investigate "Privacy-Preserving Computation of Disease Risk by Using Genomic, Clinical, and Environmental Data". It describes privacy for storing and processing unit in system. It uses homomorphism encryption and Privacy-Preserving integer comparison. It uses real patient's data and reliable disease risk factor. It works efficiently only for genomic data. It specify disease risk test using genomic data.[6]

G. *Multilink Constrained K-means algorithm*

The multilink constrained K-means algorithm assigns each point to the cluster whose centre also called centroid. The centre is the average of all the points in the cluster is coordinates the arithmetic mean. It is one of the methods used to partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining. It gives more accurate result for prediction of disease[8]

Proposed Methodology

In the proposed system the efforts are made to enhance the accuracy and performance of the system.

A. *System Architecture*

The work flow of the electricity forecasting system is represented diagrammatically in the Fig 1 given below:

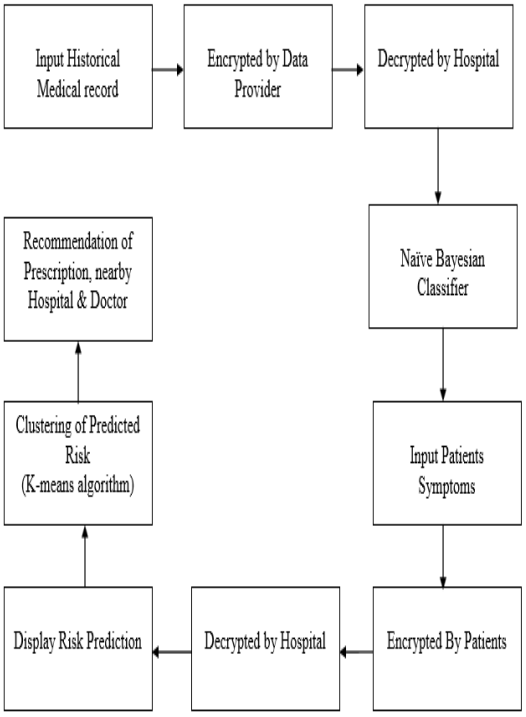

Fig 1: Proposed System Architecture for prediction of disease prevention

III. RESULT AND DISCUSSION

All patient's data is stored on cloud for privacy purpose. Naïve Bayesian gives the classification of data whereas K-means does the clustering of symptoms. Using both techniques we can find accurate prediction within less time.

Following graph shows the time taken by system to classify the disease based on symptoms, test result and historical data of patient. Using K-means clustering algorithm it will give diagnosis result in less time with high accuracy value.

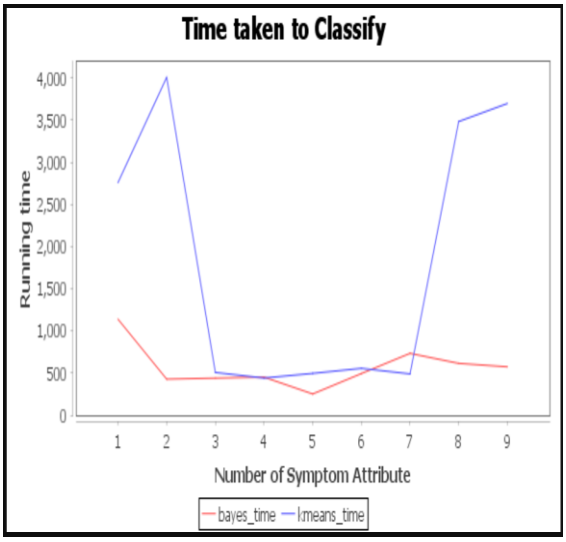

Fig 4: Graph of no of Symptom and time
(Red line: Bayes Time, Blue line: K-means time)

IV. CONCLUSIONS

The proposed system is designed to predict the prevention of disease prediction demand. Classification of disease is mainly based on patient symptoms and test result and after prediction of disease clustering of that disease is done. To do this clustering in between disease naive Bayesian classifier has many challenges. The identified challenge of clustering the disease can be resolved in future using proposed clustering method. The proposed K-means algorithm for clinical decision support system is reduce disease prediction time and increase diagnosis accuracy and can give better result for clinician

ACKNOWLEDGMENT

For everything achieved, the credit goes to all those who had really helped us to complete this work successfully. We are extremely thankful to P. G. Coordinator Prof. N.P. Kulkarni and Dr. V. S. Bidve project guide for guidance and review of this paper. I am very much grateful to our Project coordinator Dr. L. V. Patil for providing all the facilities. I would also indebted to Dr. K. R. Borole, Vice-Principal Smt. Kashibai Navale College of Engineering, Pune for providing the facilities needed for completion of the dissertation and necessary guidance. I would also indebted to Dr. A. V. Deshpande, Principal Smt. Kashibai Navale College of Engineering, Pune for providing the facilities needed for completion of the dissertation and necessary guidance. I would also like to thanks the all faculty members of "SKN College Of Engineering".

REFERENCES

[1] R. Bellazzi and B. Zupan, "Predictive data mining in clinical medicine: current issues and guidelines," International journal of medical informatics, vol. 77, no. 2, pp. 81–97, 2008.

[2] Y. Rahulamathavan, S. Veluru, R. Phan, J. Chambers, and M. Rajarajan, "Privacy-preserving clinical decision support system using Gaussian kernel based classification," IEEE Journal of Biomedical and Health Informatics, pp. 56–66, 2014.

[3] Y. Tong, J. Sun, S. S. M. Chow, and P. Li, "Cloud-assisted mobile-access of health data with privacy and auditability," IEEE J. Biomedical and Health Informatics, vol. 18, no. 2, pp. 419– 429, 2014.

[4] R. Lu, H. Zhu, X. Liu, J. K. Liu, and J. Shao, "Toward efficient and privacy-preserving computing in big data era," IEEE Network, vol. 28, no. 4, pp. 46–50, 2014.

[5] H. Monkaresi, R. A. Calvo, and H. Yan, "A machine learning approach to improve contactless heart rate monitoring using a webcam," IEEE J. Biomedical and Health Informatics, vol. 18, no. 4, pp. 1153–1160, 2014.

[6] J.-P. Hubaux, J. Fellay, E. Ayday, M. Laren, J. L. Raisaro, P. Jack et al., "Privacy-preserving computation of disease risk by using genomic, clinical, and environmental data," in Proceedings of USENIX Security Workshop on Health Information Technologies (HealthTech" 13), no. EPFL-CONF-187118, 2013.

[7] Ximeng Liu, Student Member, IEEE, Rongxing Lu, Member, IEEE, Jianfeng Ma, Member, IEEE, Le Chen, and Baodong Qin " Privacy-Preserving Patient-Centric Clinical Decision Support System on Na¨ıve Bayesian Classification ", IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS, VOL. XX, NO. XX, DECEMBER2016.

[8] M.Parvathavarthini , E.Ramara "Multilink Constrained k-means Clustering Algorithm for Information Retrieval", International Journal of Engineering Trends and Technology (IJETT), V13(3),140-143 July 2014. ISSN:2231-5381