Privacy-Preserving Clinical Decision Support System

N.Bhavani[1], R.Aarthi[2], V.Preethi[3], P.Priyadarshini[4], S.Priyanga[5]
[1, 2, 3, 4, 5]Dept of Information Technology
[1, 2, 3, 4, 5] Saranathan college of engineering, Tiruchirapalli-620012, TamilNadu, India

Abstract- *Clinical decision support system, which uses advanced datamining techniques to help clinician make proper decisions, has received considerable attention recently. The advantages of clinical decision support system include improving diagnosis accuracy and reducing diagnosis time. Specifically, with large amounts of clinical data generated every day, naïve Bayesian classification can be utilized to excavate valuable information to improve a clinical decision support system. Although the clinical decision support system is quite promising, the flourish of the system still faces many challenges including information security and privacy concerns. In this paper, we propose a privacy-preserving clinical decision support system, which helps clinician to diagnose the disease of patients' in a privacy-preserving way. In the proposed system, the past patients' historical data are stored in cloud and can be used to train the naive Bayesian classifier without leaking any individual patient medical data. It consists of health information of all the patients. This mainly helps to train the Naïve Bayesian Classifier. The new patients are the ones who have not diagnosed yet. These patients will give their health information. This information will be collected during doctor visit. The new patient can give blood pressure, sugar level, cholesterol, ECG results etc. to the Processing Unit for the diagnosis. To provide privacy and to avoid leakage in naïve Bayesian classifier, a cryptographic tool called Additive homomorphic proxy aggregation scheme is used. Privacy analysis ensures that patient's information is private and will not be leaked out during the disease diagnosis phase. In addition, performance evaluation can demonstrate that our system can efficiently calculate patient's disease risk with high accuracy in a privacy-preserving way.*

Keywords- Clinical decision support System (CDSS), naive Bayesian classifier, privacy preserving.

I. INTRODUCTION

Healthcare industry extensively distributed in the global scope to provide health services for patients, has never faced such a massive amount of electronic data or experienced such a sharp growth rate of data today. To speed up the diagnosis time and improve the diagnosis accuracy, a new system in the health care industry should be workable to provide a much cheaper and faster way for diagnosis. A clinical decision support system (CDSS) is an application that analyses data to help healthcare providers for making clinical decisions. Clinicians use a CDSS to diagnose the disease. CDSS encompasses a variety of data mining tools to enhance decision making in the clinical workflow. These tools provide clinical guidelines, diagnostic support to the patients. The advantages of clinical decision support system include not only improving diagnosis accuracy but also reducing diagnosis time. The benefits of CDSS also include increased quality of care and enhanced health outcomes, avoidance of errors and adverse events, improved efficiency, cost-benefit, and provider and patient satisfaction.

The growth of electronic health records, as well as the challenge of managing patients with complex conditions, has justified the need for improving the existing system. The main problem in the existing system is to keep patient's medical data away from unauthorised disclosure. To address the privacy issues lying in the CDSS, a privacy preserving clinical decision support system called PPCD, which is based on Naïve Bayesian classification to help physicians to predict disease risks of patients in a privacy preserving way. For minimizing patient's privacy disclosure a new aggregation technique called additive homomorphic aggregation (AHPA) scheme, which allows service provider to build naïve Bayesian classifier without leaking individual historical medical data.

II. LITERATURE SURVEY

1. Computer-assisted decision support for the diagnosis and treatment of infectious diseases in intensive care units

Diagnosing infections in critically ill patients admitted to intensive care units (ICUs) is a challenge because signs and symptoms are usually non-specific for a particular infection. In addition, the choice of treatment, or the decision not to treat, can be difficult. Models and computer-based decision-support systems have been developed to assist ICU physicians in the management of infectious diseases. We should discuss the historical development, possibilities, and limitations of various computer-based decision-support models for infectious diseases, with special emphasis on Bayesian approaches. Although Bayesian decision-support systems are potentially useful for medical decision making in infectious disease management, clinical experience with them is limited and prospective evaluation is needed to determine whether their use can improve the quality of patient care.

Decision-support systems based on models of expert knowledge are called expert systems. We review the development of such models applied to infectious disease management. We use the diagnosis of ventilator-associated pneumonia (VAP) to illustrate the differences and similarities between the various methods, in particular regarding their clinical use.

Clinicians are generally reluctant to use computerised guidelines that require additional data entry and time and effort. As a result, decision support systems still lack clinical credibility. Yet, a few decision-support systems have actually been shown to improve the care process. Therefore, as for any other new diagnostic technique or therapy, prospective trials are needed to provide the necessary evidence on which implementation and wide-scale use of decision-support systems can be based.

2. Naïve Bayesian Classifier

A statistical classifier called Naive Bayesian classifier is discussed. This classifier is based on the Bayes Theorem and the maximum posteriori hypothesis. The naive assumption of class conditional independence is often made to reduce the computational cost.

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classifier is based on Bayes theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered naive.
Bayes Theorem;

$$P(H|X) = \frac{P(X|H)\,P(H)}{P(X)}$$

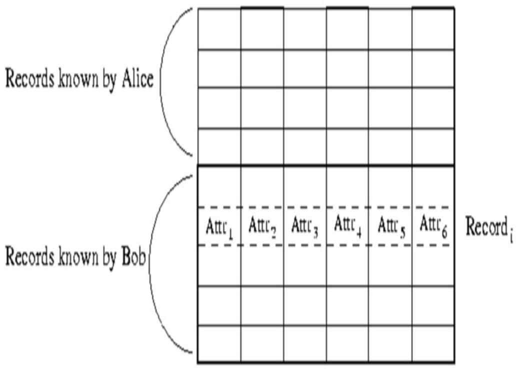
Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. All Naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature given the class variable.

The remarks on the Naive Bayesian classifiers are Studies comparing classification algorithms have found that the naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers and Bayesian classifiers have also exhibited high accuracy and speed when applied to large databases.

The advantage on the Naive Bayesian is it requires a small number of training data to estimate the parameters necessary for classification.

3. A New Efficient Privacy-Preserving Scalar Product Protocol

Privacy issues have become important in data analysis, especially when data is horizontally partitioned over several parties. In data mining, the data is typically represented as attribute-vectors and, for many applications; the scalar (dot) product is one of the fundamental operations that is repeatedly used. In privacy-preserving data mining, data is distributed across several parties. The efficiency of secure scalar products is important, not only because they can cause overhead in communication cost, but dot product operations also serve as one of the basic building blocks for many other secure protocols. Although several solutions exist in the relevant literature for this problem, the need for more efficient and more practical solutions still remains. Thus a very efficient and very practical secure scalar product protocol is presented. This is compared with the most common scalar product protocols. Thus this protocol is shown more efficient than the existing ones, and also provided with some experimental results by using a real life dataset.



Homomorphic encryption:

An encryption scheme is called additive homomorphic if it has the following property:

$$E(x1) \times E(x2) = E(x1 + x2).$$

Add Vectors Protocol:

The technique was introduced for manipulation of vector operations as the permutation protocol and is also known as the permutation algorithm. In this protocol, Alice has a vector ~x while Bob has vector ~y and a permutation π. The goal is for Alice to obtain π(~x+~y); that is Alice obtains the sum ~s of the vectors in some sense. The entries are randomly permuted, so Alice cannot perform ~s −~x to find ~y. Also, Bob is not to learn ~x. Solutions for P=2, that is, two parties, are based on homomorphic encryption.

New Privacy-Preserving Scalar Product Protocol:

A new simple scalar vector product protocol which is based on the Add Vectors Protocol is proposed. This protocol is very simple and it is easy to implement. Depending on the domain and encryption it is also very efficient.

1. *Alice and Bob apply the* ADD VECTORS PROTOCOL *for Alice to obtain $\pi_0(\vec{a}-\vec{b})$, were π_0 is a permutation generated by Bob.*
2. *Alice can obtain*

$$\sum_{\pi_0(i)=1}^{n}(a_{\pi_0(i)}-b_{\pi_0(i)})^2+\sum_{i=1}^{n}a_i^2$$

and Bob can compute $\sum_{i=1}^{n}b_i^2$. [2]

3. *Now Bob can send $\sum_{i=1}^{n}b_i^2$ to Alice which will allow Alice to compute the scalar product, that is*

$$2\vec{a}^T\cdot\vec{b}=2\sum_{i=1}^{n}a_ib_i$$
$$=\sum_{i=1}^{n}a_i^2+\sum_{i=1}^{n}b_i^2-\sum_{\pi_0(i)=1}^{n}(a_{\pi_0(i)}-b_{\pi_0(i)})^2.$$

The encryption used in the Add Vectors Protocol can be as simple as adding a random vector consisting of the same random number. This protocol is very efficient and also avoids problems in the add vector protocol by adding a small overhead in communication and computation costs.

III. RELATED WORKS

Data Mining:

Data mining refers to extracting or “mining” knowledge from large amount of data. It is considered as a synonym for another popularly used term Knowledge Discovery in Databases or KDD. It is a field at the intersection of computer science and statistics. It utilizes methods at the intersection of artificial intelligence, machine learning, statistics, and systems. The overall goal of the data mining process is to extract information from a dataset and transform it into an understandable structure. It involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post processing of discover structures, visualization and online updating.

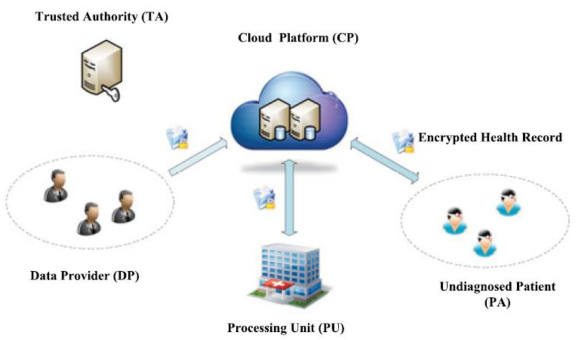
Clinical Decision Support System:

A clinical decision-support system is any computer program designed to help healthcare professionals to make clinical decisions. In a sense, any computer system that deals with clinical data or knowledge is intended to provide decision support. It is accordingly useful to consider three types of decision-support functions, ranging from generalized to patient specific.

Naïve Bayesian Classifier:

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities, such as the probability that a given sample belongs to a particular class. Bayesian classifier is based on Bayes’ theorem. Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence. It is made to simplify the computation involved and, in this sense, is considered “naive”.

IV. SYSTEM ARCHITECTURE



This system consists of processing unit that processes the clinical text. The processing unit is considered to be a hospital or a company. It helps in providing online direct-to-consumer service by providing individual risk prediction for the diseases based on the health information of a new patient. This unit consists of the historical medical data that helps to train the naïve Bayesian classifier and then finds the disease risk of the undiagnosed patients. The Data Provider provides the historical data of the old patients. It consists of health information and diseases of all the patients. This mainly helps to train the Naïve Bayesian Classifier. All these data are outsourced to the cloud platform. The new patients will give their health information such as Sugar level, blood pressure, ECG results etc. This information has to be send to the Processing Unit for the diagnosis. To provide privacy and to avoid leakage in naïve Bayesian classifier a cryptographic tool called Additive homomorphic proxy aggregation scheme is used. The trusted authority is the one that can be trusted by all other entities in the system. It distributes and manages the private keys involved in the process.

V. PROPOSED SYSTEM

To support the already existing clinical decision support system, a new privacy preserving clinical decision support system is proposed. The proposed PPCD system uses certain techniques to overcome the disadvantages in the existing CDSS.

a) TRAINING NAÏVE BAYESIAN:

The past historical data stored in the cloud is loaded as a training dataset and it is used to train the Naïve Bayesian classifier. The statistical Naïve Bayesian classifier will predict the disease risk using probabilistic values generated, while training the dataset provided by the data provider. The probabilistic value is derived using Bayes theorem.

b) PRIVACY PRESERVING:

To avoid leakage of patient information privacy can be provided using additive homomorphism. The additive homomorphism can be achieved using paillier encryption.

PAILLIER ENCRYPTION:

I) Steps for Key Generation:

- Choose two large prime numbers p and q randomly and independently of each other.
- Compute the modulus n ; the product of two primes n = p.q and λ = lcm (p − 1, q − 1) (λ is Carmichaels function)
- Select random integer g where $g \in Z^{*}_{n}{}^{2}$ and g's order is a non-zero multiple of n (since g=(1+n) works and is easily calculated - this is the best choice).
- Ensure n divides the order of g by checking the existence of the following modular multiplicative inverse: μ=L ($g\lambda$ mod n^2) $^{-1}$ mod n, where function L is defined as (Lagrange function) L (u)=u-1\ n for u =1 mod n. The public (encryption) key is (n, g). The private (decryption) key is (λ).

II) Steps for Encryption:

- Plaintext is m where m < n. $r \in Z^{*}_{n}{}^{2}$
- Find a random r
- Let cipher text c = g^m . r^n mod n^2.

III) Steps for Decryption:

- The cipher text c < n^2.
- Retrieve plaintext by calculating, m=L(cλmod n^2)/L(gλmod n^2) mod n.

VI. CONCLUSION AND FUTURE WORK

A PPCD using naive Bayesian classifier is proposed. By taking the advantage of emerging cloud computing technique, we can use big medical dataset stored in cloud platform to train naive Bayesian classifier, and then apply the classifier for disease diagnosis without compromising the privacy. In addition, the patient can securely retrieve the results of disease prediction 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. For the future work, this patient centric clinical decision support system with other advanced data mining techniques will be exploited.

REFERENCES

[1] Ms. Meenal V. Deshmukh, Prof. Swapnil N. Sawalkar, “A Survey on Privacy Preserving Data Mining Techniques for Clinical Decision Support System” in International Research Journal Of Engineering And Technology (IRJET) Volume: 03 ISSUE: 05 | MAY-2016.

[2] Andreas Steffen, “The Paillier Cryptosysem” in Institute of Internet Technologies and Applications.

[3] Eta S. Berner, “Clinical Decision Support Systems” in Second Edition of Health Informatics Series.

[4] Mark A. Musen, Yuval Shahar, And Edward H. Shortiffe, “Clinical Decision-Support Systems” .

[5] I. Rish, “An empirical study of the naive Bayes classifier” in T.J. Watson Research Center.

[6] Shruti B Karki, Mrs. Anitha K, Mrs. Nandini G, “Privacy Preserving Top-K Disease Names Retrieval Method for Clinical Decision Support System” in International

Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 4, Issue 5, May 2017.

[7] Jiawei Han, Micheline Kamber, Jian Pei, “Data Mining Concepts and Techniques” Third Edition.