A Survey on Privacy Preserving Data Mining Techniques for Clinical Decision Support System

Ms. Meenal V. Deshmukh
Student of Master of Engineering in (CSE),
Sipna College of Engineering and Technology, Amravati, India

Prof. Pritish A. Tijare
Associate professor
Sipna College of Engineering and Technology, Amravati, India

Prof. Swapnil N. Sawalkar
Assistant professor
Sipna College of Engineering and Technology, Amravati, India

----------***----------

Abstract - Clinical Decision Support System (CDSS), with various data mining techniques being applied to assist physicians in diagnosing patient disease with similar symptoms, has received a great attention recently. The advantages of clinical decision support system include not only improving diagnosis accuracy but also reducing diagnosis time. In this paper, we have given the CDSS with some advance technologies like Support Vector Machine (SVM) classifier has offered many advantages over the traditional healthcare systems and opens a new way for clinicians to predict patient's diseases. As healthcare is the field in which Security of data related to patient diseases are needs to be more secure, for that in this paper, we have use RSA and Holomorphic encryption technique that properly meets the security Goals. Specifically, with large amounts of clinical data generated every day, support vector machine (SVM) classification can be utilized to excavate valuable information to improve clinical decision support system. Although clinical decision support system is quite promising, the flourish of the system still faces many challenges including information security and privacy concerns. In this we will use homomorphic encryption technique to preserve the patient's privacy on the cloud. The patient's data can be compromised over the cloud. To overcome this scenario homomorphic encryption technique helps. The processing is done on the encrypted data; hence there is no chance of compromising privacy of patient's data.

Keywords: Clinical Decision Support System, Privacy Preserving, Support Vector Machine, Homomorphic Encryption.

1. Introduction

Now a days, Healthcare industry is extensively distributed in the global scope to provide health services for patients, has never faced such a massive amounts of electronic data or experienced such a sharp growth rate of data today. However, if no appropriate technique is developed to find great potential economic values from big healthcare data, these data might not only become meaningless but also requires a large amount of space to store and manage. Over the past two decades, the miraculous evolution of data mining technique has imposed a major impact on the revolution of human's lifestyle by predicting behaviors and future trends on everything which can convert stored data into meaningful information. These techniques are well suitable for providing decision support in the healthcare setting. 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 [1]. Clinical Decision Support System (CDSS), with various data mining techniques being applied to assist physicians in diagnosing patient diseases with similar symptoms, has received a great attention recently.

Clinical decision support system has been defined as an "active knowledge systems", which use two or more items of patient's data to generate case specific advice. This implies that a CDSS is simply a decision support system that is focused on using knowledge management in such a way to achieve clinical advice for patient care based on multiple items of patient's data. The main purpose of modern CDSS is to assist clinicians at the point of care. This means that clinicians interact with a CDSS to help to analyses, and reach a diagnosis based on patient data. Naive Bayesian classifier, one of the popular machine learning tool of data mining, has been widely used recently to predict various diseases in CDSS [1]. Despite its simplicity, it is more appropriate for medical diagnosis in healthcare than some sophisticated techniques. The CDSS with naive Bayesian classifier has offered many advantages over the traditional healthcare systems and opens a new way for clinicians to predict patient's diseases [2].

However, its flourish still hinges on understanding and managing the information security and privacy challenges, especially during the patient disease decision phase. One of the main challenges is how to keep patient's medical data away from unauthorized disclosure. The usage of medical data can be of interest for a large variety

of healthcare stakeholders. For example, an online direct-to-consumer service provider offers individual risk prediction for patient's disease. Without good protection of patient's medical data, patient may feel afraid that his medical data will be leaked and abused, and refuse to provide his medical data to CDSS for diagnosis. Therefore, it is crucial to protect patient's medical data. However, keeping medical data's privacy is not sufficient to push forward the whole CDSS into flourish. Service provider's classifier, which is used to predict patient's disease, cannot be exposed to third parties since the classifier is considered as service provider's own asset. Otherwise, the third parties can abuse the classifier to predict patient's disease which could damage service provider's profit. Therefore, besides preserving the privacy of patient's medical data, how to protect service provider's privacy is also crucial for the CDSS. Preserving Patient-Centric Clinical Decision Support System called PPCD, helps physician to predict disease risks of patients in privacy-preserving way. A secure and privacy-preserving patient-centric clinical decision support system which allows service provider to diagnose patient's disease without leaking any patient's medical data. CDSS provides knowledge and specific information about the diseases for clinicians to enhance diagnostic efficiency and improving healthcare quality. It can highly elevate patient safety, improve healthcare quality.

With increasing amounts of data being generated by all the healthcare industries and researchers there is a need for fast, accurate and robust algorithms for data analysis [8]. Improvements in databases technology, computing performance and artificial intelligence have contributed to the development of intelligent data analysis. The primary aim of data mining is to discover patterns in the data that lead to better understanding of the data generating process and to useful predictions. One recent technique that has been developed to address these issues is the support vector machine. The support vector machine has been developed as robust tool for classification and regression in noisy, complex domains.

In this paper, we have use the Support Vector Machine (SVM) technique which is one of the most powerful classification techniques that was successfully applied to many real world problems. This SVM has greater advantage in case of improving diagnosis accuracy in clinical decision support system. Along with this, we have uses the homomorphic Encryption technique for providing security to the sensitive data related to patient health information. The remaining paper is organized as, Section II gives some Literature Survey which gives some brief information of the study done in the field of Clinical Decision Support System CDSS. Section III discuss in brief about some the data mining and Privacy preserving technique used for providing CDSS. Section IV provides the workflow of the system that we have generated for providing proper CDSS. Finally, Section V concludes the paper.

2. Literature Survey

The authors, Ximeng Liu, Rongxing Lu, Jianfeng Ma in [1] proposed a privacy-preserving patient-centric clinical decision support system using naïve Bayesian classifier. By taking the advantage of emerging cloud computing technique, processing unit can use big medical dataset stored in cloud platform to train naïve Bayesian classifier. And then apply the classifier for disease diagnosis without compromising the privacy of data provider.

The authors, R. S. Ledley and L. B. Lusted [2] computer-assisted clinical decision support systems, who found that physicians have an imperfect knowledge of how they solve diagnostic problems. This article dealt with Bayesian and decision-analytic diagnostic systems and experimental proto- types appeared within a few years.

The authors [3] have performed some experiments for tumor detection in digital mammography. In this paper different data mining techniques, neural networks and association rule mining, have been used for anomaly detection and classification. From the experimental results it is clear that the two approaches performed well, obtaining a classification accuracy reaching over 70% percent for both techniques. The experiments conducted, demonstrate the use and effectiveness of association rule mining in image categorization.

The author Schurink et al. [4] discus the computer-based decision-support systems to assist Intensive Care Unit (ICU) physicians in the management of infectious diseases. In this paper, they described several computer models (such as bayesian networks) that may be used in clinical practice in the near future. As the privacy of the patient's information becomes more and more important, naive Bayesian classification were considered as a challenge to privacy-preservation due to their natural tendency to use sensitive information about individuals.

Chuang et. al [5] revealed the means of effectively using a number of validation sets obtained from the original training data to improve the performance of a classifier. The proposed validation boosting algorithm was illustrated with a support vector machine (SVM) in Lymphography classification. A number of runs with the algorithm was generated to show its robustness as well as to generate consensus results. At each run, a number of validation datasets were generated by randomly picking a portion of the original training dataset. At each iteration, the trained classifier was used to classify the current validation dataset. Experimental results on the Lymphography dataset showed that the proposed method with validation boosting could achieve much better generalization performance (on repeated iterations) with a testing set than the case without validation boosting.

Mc.Sherry et.al, [6] presented an algorithm for Conversational Case Based Reasoning (CCBR) called iNN(k) in which feature selection was motivated with the goal to confirm a target class and informed by a measure of feature's discriminating power of supporting the target class. The performance of iNN (k) on a given dataset was

shown to depend on the value of 'k' and on local or global selection of feature was used in the algorithm. Only 42% and 51% on an average of features that was needed in complete problem description by iNN (k) to provide accuracy levels of 86.5% and 84.3% respectively on the Lymphography and SPECT heart datasets from the UCI machine learning repository.

In [7], Yi et al. improved the by both efficiency and privacy and this scheme could prevent eavesdropping attack. This two-party protocol can also be easily extended to multi-party protocol. Different from horizontal partition, another kind of data partition called vertically partition (one patient's different attributes are partitioned) were introduced to privacy- preserving naive Bayesian classifier by using secure scalar product protocol [9].

3. Technologies Used for Designing

3.1 Data mining techniques

Data mining techniques have been widely used in clinical decision support systems (CDSS) that performs prediction and diagnosis of various diseases with better accuracy. The techniques have been very effective and helps in developing clinical support systems because they are able to detect hidden patterns and relationships in medical data. There are large no. of classification techniques which can be used for clinical decision support system. The aim of classification is to predict the target class for each case in the data accurately. Classification is important when a repository of data contains samples that can be used as the basis for future decision making. Some of the data mining that are mainly used for classification in CDSS are given below:

A. Bayesian Belief Network

The Bayesian network is used as knowledge based graphical representation that shows a set of variables and their probabilistic relationships between diseases and their symptoms. Bayesian network is utilized to find the probability of the presence of possible diseases given their symptoms. The advantage is that it requires the knowledge and conclusions of experts in the form of probabilities. It is very important for any physician who has no computer expertise to understand about the Bayesian network. Which is gives as a clinician reference with a searchable database of diseases and clinical manifestations. It also applies a statistical pattern-matching approach that considers the age of onset and offset of the findings in each disease [10].

B. Neural Network

Neural Networks is allows the systems to learn from existing knowledge and experiences. The three main layers of Neural Networks are Input, Output and Hidden layer. Neural Network is made of nodes that is called neurons. And there is weighted connection between nodes of different layers, which is used to transfer signals between the nodes. Neural Network is can continue with incomplete data that gives educated guesses about missing data and get improved with every use due to its adaptive system learning. Mr. P. A. Kharat et al 2011 proposed clinical decision support system DSS based on Jordan/Elman neural network for diagnosis of epilepsy and they got relatively high overall accuracy for training data is 99.83% and for cross-validation data and testing data is99.92% [11].

C. Decision Tree

Decision tree is also one of the most often used techniques of data analysis. It is applied to classify records to a proper class. In medical field decision trees is useful in determining the sequence of attributes. First it makes a set of solved cases. Then the whole set is divided into sets namely training set and testing set. A training set is used for the induction of a decision tree. A testing set is used for finding out the accuracy of an obtained solution. AY AI-Hyari et al 2013 developed a CDSS for diagnosing patients with Chronic Renal Failure using different classification methods like neural network, naïve bays and decision tree. They proved that there is (92.2%) accuracy of using Decision tree algorithm as compared to all other algorithms/ implementations involved in their study [12]. They applied supervised decision tree classifier C4.5 to classify image samples with sensitivity of 98.1% and specificity of 99.6%.

D. Naïve Bays

Naïve Bays uses the kernel estimator for numeric attributes rather than a normal distribution and that utilized Supervised Discretization while converting numeric attributes to normal ones. We got an Output in text form of Naïve Bayes classifier. [13] Mrs. G. Subbalakshmi et al 2011 developed a CDSS for heart disease prediction. Thses system extracts hidden knowledge using a historical heart disease database. They claimed that it is the most effective model for predict patients suffering with heart disease. Advantages of Naïve bays are that it is simple and efficient and it gives better performance for classification.

E. Support Vector Machine

Support vector machine (SVM) has become more and more popular tool in task of machine learning involving classification, regression etc. SVM separate the data into two categories and performing classification and then constructing an N-dimensional hyper plane. SVM is supervised the learning model applied mainly for classification. SVM serves as the linear separator between two data points to identify two different classes for multidimensional environment. SVM algorithms are in the binary format. In multi-class problem one must reduce the problem to a set of multiple binary classification problems.

They applied rough set for the purpose of feature selection and SVM for classification. They attained very high classification accuracy of 99.41% for 50–50% of training-test partition, 100% for 70–30% of training test partition, and 100% for 80 –20% of training - test partition. They were also able for discovering a combination of five informative features, which can be important to the physicians for breast diagnosis [14].

Support Vector Machine is a state of the art classification. It performs well with real world application such as classifying text, classifying images etc. SVM are the standard tools for machine learning and data mining. And with this large amount of application and advantage, we will also use SVM for our proposed classification technique in CDSS.

3.2 Privacy Requirements

Privacy is crucial for the success of patient's diseases diagnosis. In our privacy model, we consider DP is trustable which provides correct historical medical data. The internal party PU is considered as curious-but-honest which is interested in DP's individual historical medical data and PA's medical data, but strictly follows the protocols executed in the system. PA is curious about PU's classifier while CP is curious about all the other parties data in the system. Moreover, an external adversary is interested in all data transmitted in the system by eavesdropping. Therefore, in order to prevent both internal party from information leakage and external adversary from eavesdropping, the following privacy requirements should be satisfied in PPCD [1].

- DP's privacy: DP's historical medical data contain confirmed case records of patient's symptoms and confirmed diseases. These individual data contain some sensitive information which are highly related to patient's privacy. It cannot be directly exposed to untrusted parties during the transmission and storage. Otherwise, DP will not provide its own data to the other parties due to the privacy information leakage. Therefore, privacy of DP should be preserved in our system.
- PU's privacy: PU uses historical medical data to train SVM classifier and gets conditional probabilities about the classifier [14]. These probabilities are considered as an asset of PU which cannot directly be sent to patients or leaked to other parties during the disease diagnosis.
- PA's privacy: PA contains some symptom data which are sensitive and cannot directly expose to other parities. In addition, the diagnosis results are also highly sensitive information which cannot be leaked to other parties. If needed, PA can let the authorized person (authorized clinician) disclose the diagnosis results for further processing.
-

3.3 Homomophic Encryption

Homomorphic encryption is a form of encryption which allows specific types of computations to be carried out on ciphertext and obtain an encrpted result which when decrypted matches the result of operations performed on the plaintext". For example, a person can add two encrypted numbers and then the second person can decrypt the result, without being able to find the value of the individual numbers. When the data is transferred to the cloud we use standard encryption methods to secure this data, but when we want to do the calculations on data located on a remote server, it is necessary that the cloud provider has access to the raw data, and then it will decrypt them. As we all know, the demand for privacy of data and algorithms to handle the information of enterprise has increased tremendously over the last decades. To achieve this, technology such as data encryption methods with the use of tamper-resistant hardware is used. However, a critical problem arises when there is a requirement of computing on such encrypted data (publicly) where privacy is established. Hence, the homomorphic cryptosystems can be applied in this case. It is a method that enables us to perform computations on encrypted data without decryption. We will propose the application of a method to perform the operation on encrypted data without decrypting and provide the same result as well that the calculations were carried out on raw data and I will use proxy re-encryption technique that prevents ciphertext from chosen cipher text attack.

A method enable to perform the operations on encrypted data without decrypting them is homomorphic encryption. We will try to deal the problem of security of data hosted in a Cloud Computing provider [8]. We all know that the cloud or on-demand computing brings a lot of advantage to the computer science of today and tomorrow. But the adoption of Cloud passage applies only if the security is ensured. How to ensure better data security and how a client can keep their private information confidential? There are two major questions that present a challenge for providers of Cloud Computing. In this work we focus the application of Homomorphic Encryption of the security of Cloud Computing.

4. Proposed Working of System

Here, we will try to improve the existing system by using Support Vector Machine (SVM) Data Mining classification technique for Clinical Decision Support System. The system will work faster and efficient using SVM. It is widely used in real-life applications because of its simplicity and good performance both in theory and practice. We will use encryption techniques for preserving privacy of patient's data with Homomorphic Encryption Technique to re-encrypt the data on the network. All the processing will be done at server side and on the encrypted data. This achieves the privacy of patient's data and

detailed privacy analysis ensures that patient's information is private and will not be leaked out during the disease diagnosis phase. The system model of CDSS including Trusted Authority (TA), Cloud Platform (CP), Data Provider (DP), Processing Unit (PU), and Undiagnosed Patient (PA) is given in below figure 1 [1] gives the working of the system.

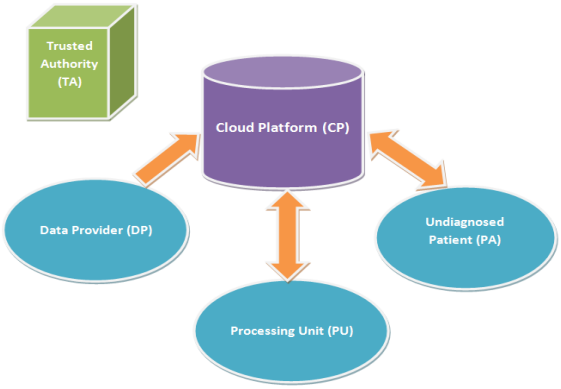

Figure 1: A System Model Under consideration [1]

1)Trusted Authority (TA): TA is the indispensable entity which is trusted by all entities involved in the system, who is in charge of distributing and managing all private keys involved in the system.

2) Cloud Platform (CP): CP contains unlimited storage space which can store and manage all the data in the system. Other parties who have limited storage space can outsourced their data to CP for storing.

3) Data Provider (DP): DP can provide historical medical data that contain patient's symptoms and confirmed diseases, which are used for training SVM classifier. All these data are outsourced to CP for storing.

4) Processing Unit (PU): PU can be a company or hospital which can provide online direct-to-customer service and offer individual risk prediction for various diseases based on client's symptoms. PU uses medical data to construct SVM classifier and then use the model to predict the disease risk of undiagnosed patients.

5) Undiagnosed Patient (PA): PA has some symptom information which is collected during doctor visits or directly provided by patient. (e.g. blood pressure, heart rate, weight, etc.). The symptoms can be sent to PU for disease diagnosis.

By designing the system like this we are able to

- Improving diagnosis accuracy for any critical diseases also
- Reducing diagnosis time gives proper prescription in much less time
- High disease prediction success rate without any kind of burden
- Reducing communication overhead
- Preserving privacy of patient's data

5. Conclusion

Clinical decision support system CDSS, which uses advanced data mining techniques that help clinician make proper decisions, has received considerable attention recently. The advantages of clinical decision support system include not only improving diagnosis accuracy but also reducing diagnosis time. We have proposed Clinical decision support system using Support Vector Machine. Using SVM, the computational time and diagnosis rate will improved. The patient can get proper much efficiently and properly by providing diagnosis result according to their own preferences. With the use of Homomorphic encryption technique, the patient's privacy over the cloud will achieved. The data that will be transfer is present in encrypted data, so that there will be no loss in the privacy of patients data while training the SVM classifier and providing data on the network.

References

[1] Ximeng Liu, Rongxing Lu, Jianfeng Ma, Le Chen, and Baodong Qin, "Privacy- Preserving Patient-Centric Clinical Decision Support System on Naïve Bayesian Classification", *IEEE JOURNAL OF BIOMEDICAL AND HEALTH INFORMATICS*, VOL. XX, NO. XX, DECEMBER 2014.

[2] R. S. Ledley and L. B. Lusted, "Reasoning foundations of medical diagnosis," Science, vol. 130, no. 3366, pp. 9–21, 1959.

[3] Maria-Luiza Antonie, Osmar R. Za¨ıane, Alexandru Coma, .Application of Data Mining Techniques for Medical Image Classification.Proceeding of second International worshop on Mutimedia data mining(MDM/KDD'2001),in conjuction with ACM SIGKDD conference.SAN FRANCISCO,USA,AUG 26,2001.

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

[5] Tzu-cheng Chuang, Okan K. Ersoy, Saul B. Gelfand, Boosting Classification Accuracy With Samples Chosen From A Validation Set, ANNIE (2007), Intelligent ` Engineering systems through artificial neural networks, St. Louis, MO, pp. 455-461.

[6] McSherry D (2011), Conversational case-based reasoning in medical decision making, Artificial Jun; 52(2):59-66. Epub

[7] X. Yi and Y. Zhang, "Privacy-preserving naive bayes classification on distributed data via semi-trusted mixers," Information Systems, vol. 34, no. 3, pp. 371–380, 2009.

[8] A. Amirbekyan and V. Estivill-Castro, "A new efficient privacy- preserving scalar product protocol," in Proceedings of the sixth Aus- tralasian conference on Data mining and analytics-Volume 70. Australian Computer Society, Inc., 2007, pp. 209–214.

[9] 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.

[10]A simultaneous consult on your patient's diagnosis, Simulconsult, www.simulconsult.com/

[11]Mr. P. A. Kharat, Dr. S. V. Dudul, IEEE 2011, Clinical Decision Support based on Jordan/Elman Network".

[12]AY AI-Hyari, A. M. Al-Taee and M. A. Al-Taee, IEEE 2013, "Clinical Decision Support for diagnosis and management of Chronic Renal Failure.

[13]G.Subbalakshmi, K. Ramesh and M. Chinna Rao,IJCSE 2011, "Decision Support in Heart Disease Prediction System using Naive Bayes"

[14]Hui-Ling Chen, Bo Yang, Jie Liu and Da-You Liu, Springer 2011, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis."