Clinical Decision Support System on SVM Classification

Abinaya G[1] abi29595@gmail.com
Dhanalakshmi M[2] dhanu111m@gmail.com
Jayanthi S[3] jayanthi.jagadeesan@gmail.com

[1]B.E Student, K.C.G College of Technology, Chennai
[2]B.E Student, K.C.G College of Technology, Chennai
[3]Assistant Professor, K.C.G College Of Technology, Chennai

***Abstract*— Clinical decision support system is a disease diagnosing tool that helps clinician to make accurate decision. As large amount of data is generated every day, Support Vector Machine, a data mining technique is used to excavate and classify the required information. The existing system encounters more risk including information security and privacy. A new clinical decision support system is proposed to help doctor to diagnose the risk of patients' disease in a privacy-preserving way. The past patients' historical health data are stored and can be used to train the SVM classifier without leaking any individual patient data. This trained classifier can be applied to compute the disease risk for new coming patients. To avoid disclosure of patients' data, a cryptographic scheme called Additive Homomorphic Proxy Aggregation [AHPA] is adopted.**

***Index Terms*—Clinical Decision Support System; Support Vector Machine; AHPA.**

I. INTRODUCTION

The development of Information Technology has generated massive amount of databases in various areas. These precious data can be manipulated using data mining techniques. Data mining holds promise in many areas of health care and medical research, with applications ranging from medical diagnosis to quality assurance. The power of data mining lies in its ability to allow users to consider data from a variety of perspectives in order to discover apparent or hidden patterns. With the advent of computing power and medical technology, large and diverse data sets and elaborate methods for data classification have been developed and studied. As a result, data mining has attracted considerable attention for the past decade, and has found its way into a large number of applications that have included both data mining and clinical decision support systems.

Decision support system refers to a computer-based system that aids the process of decision making. 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. The main purpose of modern CDSS is to assist clinicians at the point of care. SVM classifier, one of the popular machine learning tools, has been widely used recently to predict various diseases in CDSS. It has been developed as robust tool for classification and regression in noisy, complex domains. A key problem that arises in any massive collection of data is that of confidentiality. The need for privacy is sometimes due to law (e.g., for medical databases) or can be motivated by business interests. Encryption has primarily been used to prevent the disclosure of confidential information, but can also be used to provide authenticity of the source of the message. So, we propose a Clinical Decision Support System on SVM in a privacy preserving way using AHPA scheme.

II. PROPOSED SYSTEM

The system model mainly focuses on how to securely train SVM classifier and use the classifier to clinically decide patients' disease without leaking their private information. Specifically, we define the system model by dividing CDSS into four parties: Database (DB), Data Provider (DP), Processing Unit (PU), and Undiagnosed Patient (PA). The overall system model of the project can be found in figure 1.

1) Database (DB): It stores and manages all the data in the system

2) Data Provider (DP representing as administrator): DP can provide historical medical data that contain patients' symptoms and confirmed diseases, which are used for training SVM classifier. All these data are stored in the database.

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

4) 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 diseases diagnosis.

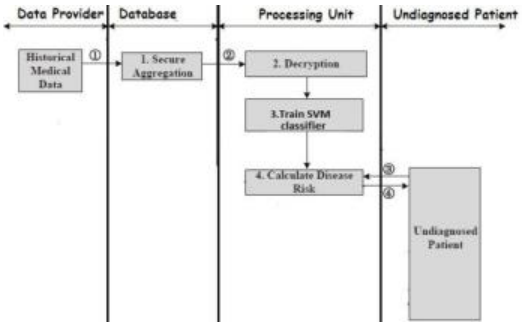


Figure 1.System Architecture

To develop the software, information of the previous patient should be given as the dataset, which is provided by the Data provider. DP will encrypt the data set and stores in the database. To provide more security to historical patients' data, the data set is re-encrypted and aggregated. When the undiagnosed or new coming patient inputs his/her symptoms, the main process takes place. Previously, the SVM classifier will be trained using the training data set given by DP. The PU will perform two stages of decryption and retrieve the data from the database. The decrypted data and the new patient's data are compared, predict the disease risk and generate the result to patient privately. This information can be viewed only by the corresponding doctor of the patient. Even the administrator of the system cannot view the medical details of the patient, he can able to view only the profile details.

Based on the system architecture, the system can be split into three modules:

i. Privacy preservation by AHPA,
ii. Training SVM classifier and
iii. Risk computation and Result generation.

Privacy preservation by AHPA

Additive Homomorphic Proxy Aggregation Scheme provides security to the patient by encrypting their details. Whenever an action occurs in the database, AHPA will be called to encrypt and decrypt the data for writing and reading. The scheme contains the following six algorithms: KeyGen, ReKeygen, Encrypt, Decrypt, Re-encrypt&Agg, and Re-decrypt. Figure 2 describes the whole process of AHPA scheme.

Step 1: KeyGen

This step will generate two public keys (Dpk, Ppk) and two private keys (Dsk, Psk) for DP and PU. The DP's keys can be used to encrypt and decrypt the patient details. Keys of PU are used to generate the re-encryption and re-decryption keys.

Step 2: ReKeygen

It is to generate the re-encryption key. This key can be generated by DP i's private key Dsk, PU's public key Ppk, and a random number 'r' is selected. Moreover, it generates a private called re-decryption key for PU to allow PU to decrypt the aggregated message using Psk and r.

Step 3: Encrypt

It is executed using DP's public key to encrypt the data for unauthorized access.

Step 4: Decrypt

The cipher text can be decrypted by using DP's private key Dsk. By decrypting, DP can view the original data.

Step 5: Re-encrypt&Agg

By using the keys generated in the ReKeygen process, the encrypted data is re-encrypted and aggregated to ensure more privacy.

Step 6: Re-decrypt

The encrypted and aggregated data are re-decrypted using the re-decryption key generated during ReKeygen process. Then the actual decryption takes place to get the original data.

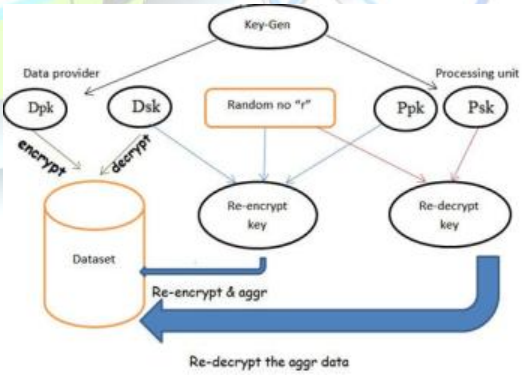


Figure 2.Additive homomorphic proxy aggregation scheme

Training SVM classifier

The support vector machine (SVM) is a widely used tool in classification problems. The SVM trains a classifier by solving an optimization problem to decide which instances of the training data set are support vectors, which are the necessarily informative instances

to form the SVM classifier. Support vectors are intact tuples taken from the training data set for classification.

Risk computation and Result generation

The trained classifier will compare the training data set and the patient data to predict the class labels i.e. it will compute the disease risk of the patient and generates the report as shown in the figure 3. The result will be sent to the patient and the doctor.

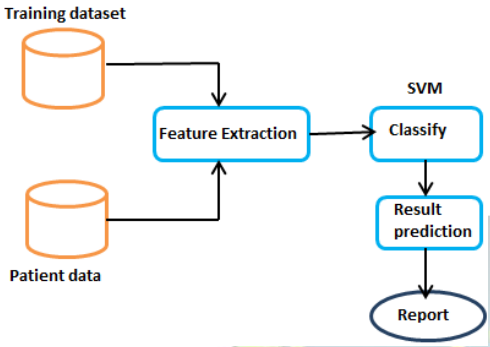


Figure 3. Result Computation

III. RELATED WORK

There is an increasing degree of concern on the privacy protection of personal information recently due to the popularity of electronic data held by commercial corporations. Data mining techniques have been viewed as a threat to the sensitive content of personal information. This kind of privacy issue has led to research for privacy preserving data mining techniques. One of the important data mining tasks is classification. As described in [2], the classification algorithm learns a classification model (i.e., the classifier) from labeled training data for the future use of classifying unseen data. There have been many privacy preserving schemes designed for various classification algorithms. The support vector machine (SVM) a powerful classification algorithm with state-of-the-art performance has also attracted lots of attention from researchers who studied privacy-preserving data mining techniques. However, a problem has still not been addressed in existing privacy-preserving SVM work[3]: the classifier learned by the SVM contains some intact instances of the training data. The classification model of the SVM inherently violates the privacy. Revealing the classifier will also reveal the private content of some individuals in the training data. Consequently the classifier learned by the SVM cannot be publicly released or be shipped to clients with privacy preservation. There is a significant difference between the SVM and other popular classification algorithms: the classifier learned by the SVM contains some intact instances of the training data. The subset of the training data kept in the SVM classifier are called support vectors, which are the informative entries making up the classifier. The support vectors are intact instances taken from the training data. The inclusion of those intact instances of the training data prevents the SVM classifier from being public releasing or shipping to client users since the release of the SVM classifier will disclose individual privacy which may violate the privacy-preservation requirements for some legal or commercial reasons. The leakage of personal information is also prohibited by laws in many countries. Most popular classification algorithms do not suffer from such direct violation of individual privacy. For example, in the decision tree classifier, each node of the decision tree stands for an attribute and denotes splitting points of the attribute values for proceeding to the next level. The Naïve Bayesian classifier [1] consists of prior probabilities of each class and class conditional independent probabilities of each value. Unlike the SVM classifier which contains some intact training instances, these classifiers merely have aggregate statistics of the training data. The paper [3] talks about decision support system for heart disease classification based on Support Vector Machine and MLP neural network architecture. We refer the paper to implement the classifier. We understand how to tackle the privacy violation problem of the classification model of the SVM, which includes some intact instances of the training data called support vectors.

IV. EXPERIMENT

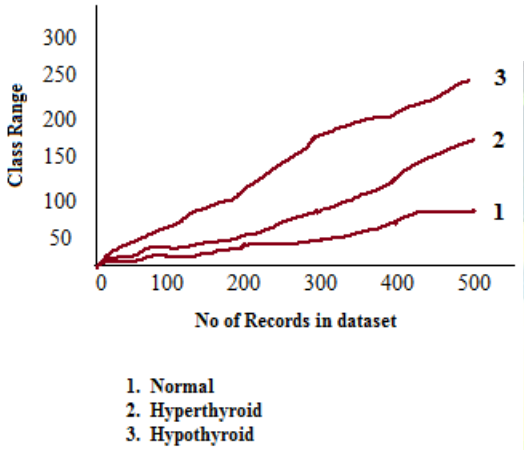
To validate the efficiency of the proposed system, a custom simulator is built in Java. It demonstrates that our can efficiently help patient to diagnose the disease with high predict success rate and it also minimizes privacy disclosure. A real dataset is used from the UCI machine learning repository called Thyroid Dataset [4]. We use this dataset to test the performance of the SVM classifier by using our Clinical Decision Support System.

Thyroid Disease data set

This data set contains several details about thyroid disease. The task is to detect, is a given patient is normal (1) or suffers from hyperthyroidism (2) or hypothyroidism (3). It is the real dataset suitable classification. It contains 7200 instances, 21 features and 3 classes. There are no missing values in the dataset. Figure 4 indicates the relation of thyroid disease.

Attribute	Domain	Attribute	Domain	Attribute
Age	[0.01, 0.97]	Thyroid_surgery	[0, 1]	Hypopituitar
Sex	[0, 1]	I131_treatment	[0, 1]	Psych
On_thyroxine	[0, 1]	Query_hypothyroid	[0, 1]	TSH
Query_on_thyroxine	[0, 1]	Query_hyperthyroid	[0, 1]	T3
On_antithyroid_medication	[0, 1]	Lithium	[0, 1]	TT4
Sick	[0, 1]	Goitre	[0, 1]	T4U
Pregnant	[0, 1]	Tumor	[0, 1]	FTI
Class	{1,2,3}			

Figure 4.Thyroid disease relation



V. CONCLUSION

Privacy preserving Clinical Decision Support System using SVM classification is proposed. The system helps the clinician to diagnose the disease risk based upon the symptoms provided by the patient. CDSS on SVM increases the accuracy of the diagnosis and reduces the diagnosis time. The system overcomes the information security and privacy challenges through Additive Homomorphic proxy aggregation scheme.

VI. REFERENCES

1. Keng-Pei Lin and Ming-Syan Chen, Fellow, (2011), "**On the Design and Analysis of the Privacy-Preserving SVM Classifier"**
2. Mrudula Gudadhe, Kapil Wankhade, Snehlata Dongre, (2010), " **Decision Support System for Heart Disease based on Support Vector Machine and Artificial Neural Network"**
3. Ximeng Liu, Rongxing Lu, Jianfeng Ma, Le Chen and Baodong Qin, **(**Dec'14**), " Privacy-Preserving Patient-Centric Clinical Decision Support System on Na¨ıve Bayesian Classification".**
4. "**Thyroid data set, UCI machine learning repository"** https://archive.ics.uci.edu/ml/datasets/Thyroid/.