ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 2, Issue 11, November 2014

Secured Neuro Genetic Approach for Predicting the Risk of Heart Disease

N.G.Bhuvaneswari Amma[1], K.Malathi[2], P.Balasubramanian[3]

Faculty, Department of Information Technology, Indian Institute of Information Technology, Srirangam, Tiruchirappalli, Tamilnadu, India[1,2,3]

ABSTRACT: Medical diagnosis is done mostly by doctors' expertise and experience. But in some circumstances, it may lead to wrong diagnosis and treatment. In this paper, a medical diagnosis system is proposed to predict the risk of heart disease using a secured neuro genetic approach. The objective of secured data classification is to build accurate classifiers without disclosing private information in the data being mined. In this paper, the learning capabilities of neural network and the optimization capabilities of genetic algorithms are combined in order to give better classification. To securely compute the activation function ElGamal scheme is used and the data is vertically partitioned. The effectiveness of the classifier is verified by experiments on Cleveland Heart Disease Dataset provided by the University of California, Irvine (UCI) machine learning repository.

KEYWORDS: Secure computing, neural networks, genetic algorithms, multiparty computation, ElGamal scheme

I. INTRODUCTION

Heart disease is a class of diseases that involve the heart, the blood vessels or both. The heart is the organ that pumps blood to all tissues of the body. If the pumping action of the heart becomes inefficient, vital organs like the brain and kidney suffer and if the heart stops working, death occurs within minutes. The World Health Organization has estimated that 12 million deaths occur worldwide, every year due to heart disease [2]. Medical diagnosis is an important yet complicated task that needs to be done accurately and efficiently. The automation of this system is very much needed to help the physicians to do better diagnosis and treatment. The representation of medical knowledge, decision making, choice and adaptation of a suitable model are some issues that a medical system should take into consideration. Medical progress is always supported by data analysis which improves the skill of medical experts and establishes the treatment technique for diseases. The purpose of medical diagnosis system is to assist physicians in defining the risk level of an individual patient. The heart disease dataset found in UCI Machine Learning Repository [22] is used for training and testing the system. The purpose of using this dataset is to provide a complex, real world data example where the relationships between the features are not easily discovered by casual inspection.

In medical diagnosis, knowledge that describes the desired system behavior is contained in datasets. When the datasets contain knowledge about the system to be designed, a neural network promises solution, because it can train itself from the datasets. The applicability of neural networks for these applications is motivated by their robustness to noisy data and their ability to determine general patterns in an efficient manner. It is possible to update a trained network by presenting new training data to the network. Using the training data, the neural network builds an internal model that maps the given data to any one of the class. Then the trained network is used to classify new data. The classification quality of the neural network depends on the number of training data. Therefore, the classification performance can be improved by using more training data. Genetic algorithm is an optimization algorithm that mimics the principles of natural genetics. It finds acceptably good solutions to problems acceptably quickly. In this paper, genetic algorithm is used to optimize the weights of the neural network.

In most of the applications, privacy issues come up because their data are considered as sensitive data. As a result, conventional methods for knowledge discovery from data are not appropriate. Therefore, secured data mining methods offer the chance to build models and extract patterns without disclosing private data. Secured data mining methods for knowledge discovery can be categorized into two groups: data perturbation methods and cryptographic methods [8] [9].

Data perturbation methods use data distortions such as adding uniform noise with the purpose of hiding private data. Cryptographic methods are used for collaborative model learning. Two or more parties contribute their data for the learning of a shared model according to security protocols that prevent the disclosure of the contributed data. In this paper, cryptographic method for supervised learning has been proposed and we focus on privacy preservation of genetic based neural network training. The data contributed among the parties are vertically partitioned in the sense that the shared model is built upon the union of the contributed datasets.

II. RELATED WORK

A number of approaches have been used to construct a secured classifier. Hai, Hussain, and Xin (2008), in their work proposed neural based learning classifier system for classifying data mining tasks [1]. Fong,Jens [20], applied a privacy preserving approach with the ID3 decision tree learning algorithm and discrete-valued attributes only. They proposed to develop the application scope for algorithms such as C4.5 and C5.0 and data mining methods with mixed discretly and continuously valued attributes. Also they proposed to optimize the storage size of the unrealized samples and the processing time when generating a decision tree from those samples.

Alka Gangrade, Ravindra Patel [20], applied a Privacy Preserving data mining method on the decision tree over horizontally Partitioned data using UTP(Un-trusted Third Party). Anand Sharma and Vibha Ojha[21], applied algorithms like ID3, Gain Ratio, Gini Index are used for constructing a decision tree. Shantakumar and Kumaraswamy (2009), in their work proposed an intelligent and effective heart attack prediction system using data mining and artificial neural network[4][5].

The first Secure Multiparty Computation (SMC) problem was described by Yao [6]. SMC allows parties with minimizing the threat of disclosure was explained [7]. Privacy preserving data mining has been an active research area for a decade. A lot of work is going on by the researcher on privacy preserving classification in distributed data mining. An overview of the new and rapidly emerging research area of privacy preserving data mining, also classify the techniques, review and evaluation of privacy preserving algorithms presented in [8]. Various tools discussed and how they can be used to solve several privacy preserving data mining problem [9]. Cryptographic research on secure distributed computation and their applications to data mining were demonstrated by Pinkas Benny [10]. Algorithm ID3 particularly a well designed and natural solution for classification was first proposed by Quinlan [11]. Lindell and Pinkas proposed a secure algorithm to build a decision tree using ID3 over horizontally partitioned data between two parties using SMC [12]. A generalized privacy preserving variant of the ID3 algorithm for vertically partitioned data distributed over two or more parties introduced in [12]. A decision tree algorithm over vertically partitioned data using secure scalar product protocol proposed in [12].A novel privacy preserving distributed decision tree learning algorithm [14], that is based on Shamir [16] and the ID3 algorithm is scalable in terms of computation and communication cost, and therefore it can be run even when there is a large number of parties involved and eliminate the need for third party and propose a new method without using third parties.

Algorithms on building decision tree, however, the tree on each party doesn't contain any information that belong to other party [15]. The drawback of this method is that the resulting class can be altered by a malicious party. Privacy preserving decision tree algorithm over vertically partitioned data, which is based on idea of passing control from site to site proposed by Wei Fang and Yang [17]. The main purpose of data classification is to build a model (i.e., classifier) to predict the (categorical) class labels of records based on a training data set where the class label of each record is given. The classifier is usually represented by classification rules, decision trees, neural networks, or mathematical formulae that can be used for classification. The work of Agrawal and Srikant [18] utilized a randomization-based perturbation approach to perturb the data. The data are individually perturbed by adding noise randomly drawn from a known distribution. A decision tree classifier is then learned from the reconstructed aggregate distributions of the perturbed data. In [19], a condensation based approach is proposed. Data are first clustered into groups, and then pseudo data are generated from those clustered groups.

Data mining tasks are then done on the generated synthetic data instead of the original data.Comparing to the works discussed above, the work discussed in this paper is different by using genetic based neural network to construct privacy preserving classifier. The activation function of the neural network is computed using the ElGamal scheme to preserve the privacy of the classifier. The neural network is used to train the system. The weights of the neural network are optimized using genetic algorithm.

III. PROPOSED ALGORITHM

The block diagram of the proposed system is illustrated in Figure 1. The major components of the system are Neural Network Owner, Training Engine, Optimization Engine, and Classification Engine.

A. Heart Disease Data Set

The Cleveland heart disease data provided by the UCI Machine Learning Repository [22] is used for analysis of this work. The dataset has 13 numeric input attributes namely age, sex, chest pain type, cholesterol, fasting blood sugar, resting ecg, maximum heart rate, exercise induced angina, old peak, slope, number of vessels colored and thal. It also has the predicted attribute ie) the class label.

B. Neural Network Owners

In this paper, we proposed a two party distributed algorithm for privacy preserving genetic
based neural network training with vertically partitioned data. We considered that there are two neural network owners, each having their own set of data. These two owners have to build one neural network based on all the data, but each party does not reveal their own data to each other. There is only one hidden layer in the neural network and the hidden nodes are chosen based on trial and error.

Elgamal is a public key encryption scheme [24] which can be defined on any cyclic group. Let G be a cyclic group with prime order q and generator g. The components of ElGamal scheme are Key generation, Encryption, and Decryption.

Algorithm 1: Elgamal Scheme

Step 1: A value x belongs to Z_p is randomly chosen as the private key. The corresponding public key
is (G,q,g,h), where $h=g_x$

Step 2: A message m belongs to G is encrypted as follows: A value r belongs to Z_p is chosen as
random. Then the ciphertext is constructed as $(C_1,C_2)=(g_r,m.h_r)$

Step 3: The plain text is computed as

$$\frac{C_2}{C_1}x = \frac{m.h_r}{g_{x}.r} = \frac{m.h_r}{h_r} = m$$

In this paper, we are considering two parties and each party only holds one part of the input to the sigmoid function. This algorithm enables them to compute the approximate value of the function without knowing the part of input from the other party. Actually, in this algorithm there is no way for each party to explore the input of the other, but the function value can still be computed. Formally, the input of the algorithm is x_1 held by party A, x_2 held by party B. The output of function y, $y(x_1 + x_2)$, is also randomly shared by the two parties. Note that the parties can always exchange their random shares of result at the end of the algorithm, so that they can learn the complete value of sigmoid function.

Algorithm 2: Sigmoid Function

Step 1: Party A generates a random number R and computes $m_i=y(x_1+i)-R$, $-n<i<=n$. Party A
encrypts each m_i using Algorithm1 and gets $E(m_i,r_i)$,where each r_i is a new random number.
It sends each $E(m_i,r_i)$ in the increasing order of i.

Step 2: Party B picks $E(m_{x2},r_{x2})$, randomizes it and sends $E(m_{x2},r_1)$ back to A, where $r_1=r_{x2}+s$, and s
is only known to party B.

Step 3: Party A partially decrypts $E(m_{x2},r_1)$ and sends partially decrypted message to B

Step 4: Party B finally decrypts the message to get $m_{x2}=y(x_1+x_2)$ - R. R is only known to A and m_{x2}
is only known to B. The function f(x) is computed as, $mx_2+R=y(x_1+x_2) = f(x)$

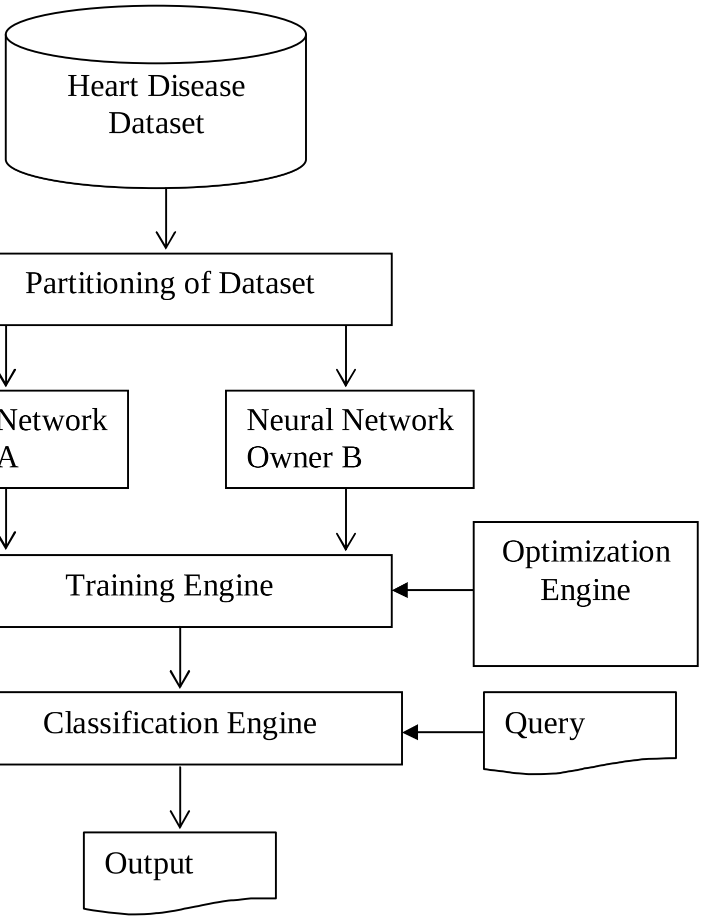

Figure 1. Block Diagram of Proposed System

C. *Training Engine*

The algorithm used to train the network is a modified version of back propagation algorithm. The input layer consists of 13 nodes, the hidden layer consists of 7 nodes and the output layer consists of 5 nodes. The hidden layer nodes are selected based on trial and error. The activation function used by the neurons is the sigmoid function [25].

D. *Optimization Engine*

Genetic algorithm is used to optimize the weights of the neural network. The proposed neuro genetic approach is the combination of training and weight optimization subsystem. The proposed algorithm is explained as follows: Let p, q, and r represents the number of neurons in the input, hidden, and output layers respectively. W_{ij} represents the weight vector for the vertices between input layer nodes to hidden layer nodes. W_{jk} represents the weight vector for the vertices between hidden layer nodes to output layer nodes. Eo represents the expected output for a given data Xo [23].

Algorithm 3: Proposed Algorithm

```
Step 1: Population of chromosomes is initialized with random numbers between 0 and 1.
Step 2: For every chromosome in the population
Step 2.1: Initialize root mean square error (RME) to 0
Step 2.2: for every data $Xo(x_1,x_2,\ldots,x_p)$, in the training set, compute
Step 2.2.1: Output of the input layer neurons
                $O_{iI}=x_i$, i=1 to p
Step 2.2.2: Input to the hidden layer neurons
```

```
            For each hidden layer neuron hj, party A computes $\sum_{j<= mA} W_{ij}x_j$ and party B
            computes $\sum_{mA< j <=mA+mB} W_{ij}x_j$
Step 2.2.3: Output of the hidden layer neurons
            Using Algorithm 2, Party A and B jointly computes the sigmoid function for each
            hidden layer node $h_j$ and obtain the random shares $h_{j1}+h_{j2}=f(\sum_j W_{ij}x_j)$
Step 2.2.4: Hidden to the output layer neurons
            For each output layer node Ok, party A computes $O_{k1}=\sum_k W_{jk}h_{j1}$ and Party B
            computes $O_{k2}=\sum_k W_{jk}h_{j2}$, such that $O_k=O_{k1}+O_{k2}$
Step 2.2.5: Output of the output layer neurons
            Using Algorithm 2, Party A and B jointly computes the sigmoid function for each of
                    the output layer neurons.
           Step 2.2.6: Cumulative error
                    $err=err+Root(O_{ko}-A_{ko})^2/r$
Step 2.3: Calculate the average error
     Avgerr=err/n
Step 2.4: Compute the fitness of the chromosome
     Fitness=1/avgerr
Step 3: If the threshold value is not reached
    Step 3.1: select top 50% of the chromosome based on the fitness value
    Step 3.2: Apply crossover and mutation on the selected chromosomes to obtain another 50% of the
              chromosomes
Step 4: goto step 2
```

E. Classification Subsystem

The trained neural network is used for classification. If the user submits a query to the classification subsystem, the privacy preserving neuro genetic classifier predicts the risk of heart disease data.

IV. SIMULATION RESULTS

The Cleveland Heart Disease Dataset provided by the UCI Machine Learning Repository [9] is used for training and testing the medical diagnosis system. The distribution of dataset is given in Table 1. Among the 303 instances of data, 200 instances are used for training and 103 instances are used for testing.

Table 1.Distribution of Data

Class	*0*	*1*	*2*	*3*	*4*
	Absent	*Low*	*Medium*	*High*	*Serious*
Training	109	38	20	23	10
Testing	55	17	16	12	3

The classifier classifies that the given data belongs to presence of heart disease or absence of heart disease. A confusion matrix contains information about actual and predicted classifications done by the classification system. The training and testing dataset classification by the proposed approach is given in Table 2. and Table 3. respectively.

Table 2.Classification of Training Data

Class	*0*	*1*	*2*	*3*	*4*
	Absent	*Low*	*Medium*	*High*	*Serious*
Yes	106	37	20	22	10
No	3	1	0	1	0

Table 3.Classification of Testing Data

Class	*0* *Absent*	*1* *Low*	*2* *Medium*	*3* *High*	*4* *Serious*
Yes	50	17	16	11	2
No	5	0	0	1	1

The performance measures of the testing data are given in Table 4. The accuracy and the precision of the classifier is calculated using the true positive, false positive, true negative and false negative values. The classification accuracy of testing set is 93.2%.

Table 4.Performance Measures

Measures	*Classifier*
True Positive	90.91%
False Positive	4.17%
True Negative	95.83%
False Negative	9.09%
Accuracy	93.2%
Precision	96.15%

V. CONCLUSION AND FUTURE WORK

In this paper, a privacy preserving neuro genetic approach is proposed to predict the risk of heart disease data. The dataset for analysis purpose is taken from UCI machine learning repository. The datasets are vertically partitioned into two sets and given to two neural network owners, each owner having their own set of data. These two owners jointly build a neural network which is securely trained using the genetic based neural network algorithm. The final network is used for classification. The classification accuracy obtained using this approach is 93.2%.

There are many interesting aspects for future work. The datasets can be partitioned horizontally to construct a classifier. More than two neural network owners can be considered. More than one hidden layer can be used to construct the classifier. The activation function can be varied in each layer.

REFERENCES

[1] Dam,H., Hussain A.Abbass Hai and Xin Yao, “Neural – Based Learning Classifier Systems”, IEEE Transactions on Knowledge and Data Engineering, Vol.20, No.1, pp.26-39, 2008.
[2] Goenka, S., Prabhakaran, D., Ajay, V.S., and Reddy, K.S., “Preventing Cardiovascular Disease in India – Translating Evidence to Action”, Current Science, Vol.97, No.3, pp.367-377, 2009.
[3] Shantakumar B.Patil and Kumaraswamy, Y.S., “Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network”, European Journal of Scientific Research, Vol.31, No.4, pp.642-656, 2009.
[4] Shantakumar B.Patil and Kumaraswamy,Y.S., “Extraction of Significant Patterns from Heart Disease Warehouses for Heart Attack Prediction”, International Journal of Computer Science and Network Security ,Vol.9, No.2, pp.228-235, 2009.
[5] Andrew C. Yao, “Protocols for secure computation,” IEEE Symposium on Foundations of Computer Science (FOCS), No.23, pp. 160-164, 1982.
[6] Wenliang Du, Mikhail J. Attalah, “Secure multi-problem computation problems and their applications: A review and open problems,” Tech. Report CERIAS Tech Report 2001-51, Center for Education and Research in Information Assurance and Security and Department of Computer Sciences, Purdue University, West Lafayette, IN 47906, 2001.
[7]Verykios, V., Bertino, E., “State-of-the-art in Privacy preserving Data Mining,” SIGMOD, Vol.33, No.1, 2004.
[8]Cliffton, C., Kantarcioglu, M., Vaidya, J., “Tools for privacy preserving distributed data mining,” ACM SIGKDD Explorations Newsletter, Vol.4, No.2, pp.28-34, 2004.
[9] Benny Pinkas, “Cryptographic techniques for privacy-preserving data mining,” ACM SIGKDD Explorations Newsletter, Vol.4, No.2, pp. 12-19, 2006.
[10]Quinlan, J.R., “Induction of decision trees,” in: Jude W. Shavlik, Thomas G. Dietterich, (Eds.), Readings in Machine Learning. Morgan Kaufmann, 1990, Vol.1, pp.81–106.
[11] Yehuda Lindell, Benny Pinkas, “Privacy preserving data mining,” Journal of Cryptology Vol.15, No.3, pp. 177–206, 2002.
[12] Vaidya, J.,Clifton, C., Kantarcioglu, M., Patterson, A. S., “Privacy-preserving decision trees over vertically partitioned data,” IFIP WG 11.3 Working Conference on Data and Applications Security, No.19, pp.139–152, 2008.

[13] Wenliang Du, Zhijun Zhan, “Building decision tree classifier on private data,” In CRPITS, pp.1–8, 2002
[14] Emekci, F.,Sahin, O.D., Agrawal, D., Abbadi, A. El., “Privacy preserving decision tree learning over multiple parties,” Data & Knowledge Engineering 63, pp. 348-361, 2007.
[15] Shamir, A., “How to share a secret,” Communications of the ACM, Vol.22, No.11, pp. 612-613, 1979.
[16] Weiwei Fang, Bingru Yang, “Privacy Preserving Decision Tree Learning Over Vertically Partitioned Data,” International Conference on Computer Science & Software Engineering, 2008.
[17] Agrawal,R., and Srikant, R., “Privacy Preserving Data Mining,” ACM SIGMOD Int’l Conf. Management of Data, 2000.
[18] Aggarwal, C.C., and Yu, P.S., “A Condensation Approach to Privacy Preserving Data Mining,” Ninth Int’l Conf. Extending Database Technology (EDBT), 2004.
[19] Pui K. Fong and Jens H.Weber-Jahnke, “Privacy preserving Decision Tree Learning Using Unrealized Data Sets”, IEEE transactions on Knowledge & Data Engineering , February 2012.
[20] Alka Gangrade, Ravindra Patel, “Privacy Preserving Two-Layer Decision Tree Classifier for Multiparty Databases”, International Journal of Computer and Information Technology, pp.2277-0764, September 2012.
[21] Anand Sharma and Vibha Ojha, “ Implementation of Cryptography for Privacy preserving data mining”, International Journal of Database Management Systems(IJDMS) Vol.2, No.3, August 2010.
[22] UCI Cleveland Heart Disease Dataset, available at http://archive.ics.edu/ml/datasets/Heart+Disease, 2009.
[23] Rajasekaran, S., and Vijayalakshmi Pai, G.A, Neural Networks, Fuzzy Logic, and Genetic Algorithms Synthesis and Applications, Prentice Hall of India, 2007.
[24] ElGamal, T., "A Public-Key Cryptosystem and a Signature Scheme Based on Discrete Logarithms, IEEE Transaction on Information Theory, Vol. IT-31, No.4, pp.469-472, 1985.
[25] Sayyad, S.S., and Kulkarni, P.J.,” Privacy Preserving Back Propagation Algorithm for Distributed Neural Network Learning”, International Journal of Scientific and Research Publications, Vol. 2, No.3, 2012.

BIOGRAPHY

N.G.Bhuvaneswari Amma is a Faculty in the Department of Information Technology, Indian Institute of Information Technology, Srirangam, Tiruchirappalli, Tamilnadu, India. She received the B.E degree in Information Technology from Jayamatha Engineering College, Aralvaimozhi, India in 2004 and the M.E degree in Computer Science and Engineering in 2009 from College of Engineering, Anna University Guindy Campus, Chennai, India. Her research interest includes Data Mining, Soft Computing and Information Security.

K. Malathi is a Faculty in the Department of Information Technology, Indian Institute of Information Technology, Srirangam, Tiruchirappalli, Tamilnadu, India. She received the B.Tech degree in Information Technology in 2006 and M.E degree in Computer Science and Engineering in 2013 from J. J. College of Engineering and Technology, Tiruchirappalli, India. Her research interest includes Image Processing, Data Mining and Computer Networks.

Balasubramanian Palani received Master of Engineering in Computer Science and Engineering from Anna University Chennai, Tamilnadu State, India and is currently teaching M.Tech Courses at Indian Institute of Information Technology (IIIT) Srirangam ,Tiruchirappalli, Tamilnadu, India . His professional interests focus on Data Mining, Information Retrieval and Web services, and his current projects include Semantic based Service Discovery, Ontology based Information Retrieval.