International Journal of Applied Engineering Research, ISSN 0973-4562 Volume 13, Number 21 (2018) pp. 15196–15202

Privacy Preserving Ant Colony Optimization based Neural Learning Classifier

N. G. Nageswari Amma
Research Scholar, Department of Computer Applications, Manonmaniam Sundaranar University, Tirunelveli, India.

Dr. F. Ramesh Dhanaseelan
Professor, Department of Computer Applications, St. Xavier's Catholic College of Engineering, Chunkankadai, India.

Abstract: Due to the advancement of technologies and usage of smart devices generate huge volumes of personal data which are more sensitive and sharing of these data is important for decision making in the competitive business world. But the knowledge discovery process creates privacy and legal issues. Therefore, privacy preserving is necessary to hide sensitive data in the knowledge discovery process. In this paper, a privacy preserving ant colony optimization based neural learning classifier algorithm is proposed, which allows to learn knowledge from the neural network without revealing the sensitive data to other parties. The objective of privacy preserving data classification is to build classifiers that non-reveal the sensitive information in the data being classified. In this paper, Backpropagation algorithm is used for classification and optimizing the weights of neural network is done by ant colony optimization. The activation function in the neural network is securely computed using ElGamal scheme and the data is vertically partitioned. Experiments were conducted to observe the effectiveness of the classifier on real world datasets downloaded from the University of California, Irvine (UCI) machine learning repository and it is evident that the proposed privacy preserving ant colony optimization based neural learning classifier is promising from the Receiver Operating Characteristics analysis.

Keywords: Ant colony optimization, backpropagation algorithm, neural network, privacy preserving.

INTRODUCTION

The popularity of electronic data creates an increasing demand in commercial corporations for the privacy protection of personal information. Data mining techniques are a threat to the sensitive content of personal data. The privacy issues have led to the research for privacy preserving data mining [1]. The main important data mining task is classification. Neural networks is one among these techniques designed for various classification. Applications, viz., smart cities, medical diagnosis, telecommunication sectors, bio-informatics, and intrusion detection, the knowledge is hidden in the datasets [2].

The neural networks provide promising solution, if the datasets contain knowledge about the system to be designed as it can be trained itself from these datasets. The application of neural networks is achieved by its main characteristics, i.e.) robustness to noisy data and its ability to determine general patterns hidden in the data in an efficient manner [3]. The updations are possible in the trained network by presenting new training data to the network. The neural network builds an internal model using the training data that maps the given training data to any one of the class label. Hence, the trained network is used to classify new instances of data. The efficiency of the neural network classification depends on the number of training data. Therefore, the performance of classification is improved

when more training instances are incorporated for training [4].

Ant Colony Optimization is the metaheuristic algorithm which is inspired by the foraging behavior of ants. This algorithm consists of two modes, viz., forward and backward modes. The forward mode construct solutions probabilistically from the population of ants based on pheromone trails. The backward mode construct solutions with quality by updating pheromone trails. The ants will converge to provide optimum solution after serveral iterations [5].

The privacy issues in most applications come up because the data in those applications are considered as sensitive data and the conventional methods for pattern discovery from data are not appropriate [6]. Therefore, privacy preserving data mining methods build models and extract patterns without disclosing private data. Privacy preservation methods for pattern discovery is classified into two groups: data perturbation methods and cryptographic methods. 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 to prevent the disclosure of the sensitive data [7].

In this paper, cryptographic method for supervised learning has been proposed and we focus on privacy preservation of ant colony optimization based neural network learning. The data are vertically partitioned among users in the sense that the shared model is built upon the union of the contributed datasets.

RELATED WORKS

The approaches to construct a privacy preserving classifier is given as follows:

Fong and Jens, in their work proposed a privacy preserving classifier with ID3 decision tree learning algorithm. The attributes considered for training are testing are with discrete valued attributes. They optimized and developed the storage size of the unrealized samples. The application considered are in the purview of C4.5 and C5.0 algorithms. The approach was applied in data mining methods with mixed discrete and continuous valued attributes also. They proved that the processing time is low when generating decision trees from the training samples [8].

Alka and Patel, in their work proposed a privacy preserving data mining method using decision tree over horizontally partitioned data with untrusted third party. The intermediate result was individually calculated for every party and it is sent to the untrusted third party. They proved that the performance of privacy preserving two-layer horizontally partitioned ID3 decision tree classifier is better than the basic ID3 decision tree classifiers [9].

Du and Zhan, in their work applied algorithms like ID3, Gain Ratio, Gini Index for constructing a decision tree. They used association rules to perform data generalization, summarization, and characterization and they faced secure multiparty communication problem. They used Oblivious Transfer Protocol for secure transmission of information. They stated that their work is better in terms of measures, viz., trust, correctness, efficiency, and fairness [10].

Quinlan proposed ID3 decision tree classification and showed that it is the best classification algorithm. Yehuda and Benny proposed a secure algorithm to build a decision tree using ID3 over horizontally partitioned data between two parties using secure multiparty computation. They also introduced a variant of general privacy preserving classifier using ID3 algorithm for vertically partitioned data which is distributed over two or more parties [11].

Vaidhya and Clifton [12] proposed a secure scalar product protocol to classify using decision tree algorithm over vertically partitioned data. A novel privacy preserving distributed decision tree learning algorithm, that is based on Shamir [13] 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. The drawback of this method is that the resulting class can be altered by a malicious party [14].

Weiwei Fang and Yang, proposed that Privacy preserving decision tree algorithm over vertically partitioned data, which is based on idea of passing control from site to site .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 [15].

Lindell and Pinkas, in their work utilized a randomization based perturbation approach to perturb the data. The data are individually perturbed by adding noise randomly drawn from a known distribution [16]. A decision tree classifier is then learned from the reconstructed aggregate distributions of the perturbed data. In the condensation based approach [17], the 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 [18].

Yaping et al., in their work proposed a privacy preserving data mining approach by enabling multilevel trust among parties. They allowed more trusted data miner to access the perturbed copy of the data and disallowed malicious data miner [19]. Keng and Ming, in their work proposed a privacy preserving Support Vector Machine (SVM) classifier. In their work, the SVM trains the classifier to decide which of the training dataset support vectors are used. The classifier designed by them violates the privacy. So they post-process the classifier to preserve the privacy of the sensitive data [20].

Bertha, David, and Santiago, in their proposed a privacy preserving distributed learning system based on genetic algorithms and artificial neural networks. Their system solved the machine learning challenges like tackling massive databases, learning in distributed environment and preserving the privacy of sensitive data [21]. Bertha et al., in their work proposed a privacy preserving distributed and incremental learning method for intrusion detection. They used artificial neural networks for incremental learning and genetic algorithms for determining relevant inputs. Mavrovouniotis and Yang proposed a pattern classification algorithm with neural network and ant colony optimization with better efficiency achievement compared to other optimization algorithms [5].

Comparing to the works discussed above, the work discussed in this paper is different by using ant colony optimization 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 backpropagation algorithm is used to train the system. The weights of the neural network are optimized using ant colony optimization.

PRIVACY PRESERVING ANT COLONY OPTIMIZATION BASED NEURAL LEARNING CLASSIFIER

The block schematic of the proposed approach is illustrated in Fig. 1. The major components of the system are Users, Backpropagation learning, Ant colony optimizer, and Classifier. The datasets provided by UCI are used for training the proposed approach and also to test the efficiency of the proposed approach. The datasets used are Iris, Diabetes, and Sonar. These datasets differ in their characteristics, in the number of features, class labels, size of the datasets, and data distributions. The neural network architectures differ for each datasets.

In this paper, a two party distributed algorithm for privacy preserving ant colony optimization based neural network training with vertically partitioned data is presented. There are two users, viz., A and B, each having their own set of data. These two users have to build one neural network based on all the data, but each user does not reveal their data to each other. The privacy preservation is performed using ElGamal scheme [22]. The ElGamal scheme is a public key encryption scheme which can be defined on any cyclic group. Let C_G be a cyclic group with prime order q and generator g. The components of ElGamal scheme are key generation, encryption, and decryption. The key generation is performed by randomly choosing a private key, Z_p. The public key is (G, q, g, h), where $h = g_x$. The encryption is performed as follows:

$$(CT_1, CT_2) = (g_r, m.h_r) \tag{1}$$

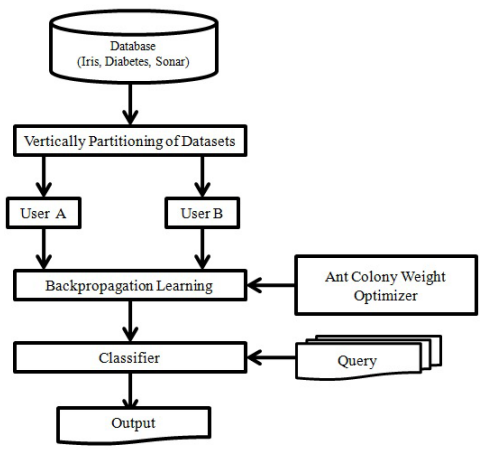

FIGURE 1. Block Schematic of Proposed Approach

where CT_1 and CT_2 are the cipher texts. The corresponding plain texts are computed as follows:

$$\frac{CT_2}{CT_1} = \frac{m.h_r}{g_x.r} = \frac{m.h_r}{h_r} = m \tag{2}$$

Two users, viz., A and B hold part of the input to compute the sigmoid function. But each user are not aware of the content what other user holds. Without revealing the sensitive information to other user, each user has to compute the sigmoid activation function. The input x_1 is held by user A and x_2 is held by other user. Both can always share their random shares at the end of the computation. The output y is computed as $y(x_1, x_2)$. By sharing the random share both the users know the value of sigmoid activation function.

The user A generates a random number, R and computes $m_i = y(x_1 + i) - R, -n < i - n$. User A encrypts each m_i using ElGamal Scheme and gets $E(m_i, r_i)$, where each r_i is a new random number. User sends each $E(m_i, r_i)$ in the increasing order of i. The User B picks $E(mx_2, rx_2)$, randomizes it and sends $E(mx_2, r_1)$ back to user A, where $r1 = rx_2 + s$, and s is only known to user B. The user A partially decrypts $E(mx_2, r_1)$ and sends partially decrypted message to user B. The user B finally decrypts the message to get $mx_2 = y(x_1 + x_2) - R$. Now R is only known to user A and mx_2 is only known to user B. The sigmoid function is computed as follows:

$$Mx_2 + R = y(x_1 + x_2) = f(x) \tag{3}$$

The algorithm used to train the network is backpropagation algorithm [23] and the weights of the neural network are optimized using Ant Colony Optimization. Let I, H, and O represents the number of neurons in the input, hidden, and output layers respectively. The weights between input to hidden layer is represented as W_{ij}. The weights between hidden to output layer is represented as W_{jk}. The expected output for a given data I_o is represented as E_o. The pheromone values are initialized with random numbers between -1 and $+1$. The root mean square error is initialized to 0. For each record in the training dataset, compute the output of input to hidden layers, and hidden layers to output. The computation in user A and user B is as follows:

$$IH_A = \sum W_{ij} \times I_o, j \leq m_A \tag{4}$$

$$IH_B = \sum W_{ij} \times I_o, m_A < j \leq m_A + m_B \tag{5}$$

The sigmoid activation function of each layer is computed jointly by user A and B and the random shares are obtained as follows:

$$H_{jA} + H_{jB} = f(\sum W_{ij} \times Ij) \tag{6}$$

The hidden to output layers are computed as follows:

$$HO_A = \sum W_{jk} \times H_{jA} \tag{7}$$

$$HO_B = \sum W_{jk} \times H_{jB} \tag{8}$$

The sigmoid function in the output layer is computed jointly by user A and user B. The root mean square error is computed as follows:

$$Error = Error + \frac{\sqrt{(E_o - A_o)^2}}{n} \tag{9}$$

where n is the number of records. The next generation of weights is computed using ant colony optimization algorithm. The pheromone value, p_{ij} is computed as follows:

$$p_{ij} = \frac{1}{\text{number of weights in neural network}} \tag{10}$$

The pheromone is updated as follows:

$$p_{ij} = p_{ij} + \Delta_{p_{ij}}^{best} \tag{11}$$

where $\Delta_{p_{ij}}^{best}$ is the amount of pheromone deposited and is computed as follows:

$$\Delta_{p_{ij}}^{best} = \frac{1}{Error^{best}} \tag{12}$$

Therefore, as the mean square error is less, more pheromone is deposited. The pheromone evaporation is to reduce the pheromone trials and is computed as follows:

$$p_{ij} = (1 - \rho)p_{ij} \tag{13}$$

where ρ is the evaporation rate and lies between 0 and 1. This weight updation procedure is repeated till the algorithm reaches the pre-defined threshold. Fig. 2 depicts the flow diagram of neural network learning. The parameters needed for ant colony optimization based neural network learning are initialized and if the threshold is not reached for the training process, the best combination of ants are selected and the learning happens. If not, the ants are released and best ants are selected and the pheromone for the best ants are updated. Likewise, the algorithm is executed for many iterations till the threshold reaches the pre-determined level.

RESULTS AND DISCUSSIONS

The datasets provided by UCI are used for training our system and also to test the efficiency of our system. The

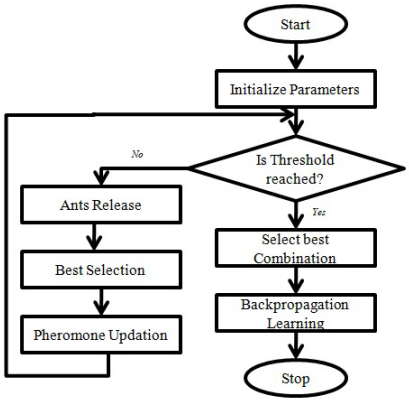

FIGURE 2. Flow Diagram of Neural Network Learning

Table 1. Datasets and Neural Network Architecture

Dataset	Number of Instances	Number of Classes	Architecture
Iris	150	3	Input Layer Nodes: 4 Hidden Layer Nodes: 4 Output Layer Nodes: 3
Diabetes	768	2	Input Layer Nodes: 8 Hidden Layer Nodes: 5 Output Layer Nodes: 1
Sonar	104	2	Input Layer Nodes: 60 Hidden Layer Nodes: 10 Output Layer Nodes: 1

datasets used are Iris, Diabetes, and Sonar. Table 1 tabulates the datasets and neural network architecture used to construct the privacy preserving genetic based neural learning classifier. There is only one hidden layer in the neural network and the hidden nodes are chosen based on trial and error. The datasets are vertically partitioned and two users privacy preserving ant colony based neural network algorithm is used training the classifier.

The distribution of UCI datasets is shown in Fig. 3. The time taken for training the network by changing the training samples is shown in Fig. 4, Fig. 5, and Fig. 6 for the datasets Iris, Diabetes, and Sonar respectively. The training size does not have significant effect on the training time. The training time is proportionate to the number of generations for which the training has been carried out. The system is evaluated by employing 10 fold cross validation using the three datasets. The network is trained through 1000 iterations for each fold. Fig. 7 depicts the error rates of the proposed algorithm for training and testing datasets. Fig. 8 shows the Receiver Operating Characterics (ROC) analysis of the proposed approach for the three datasets. It is evident that the results are promising.

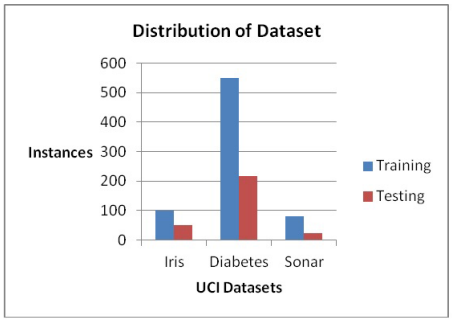

FIGURE 3. Distribution of UCI Datasets

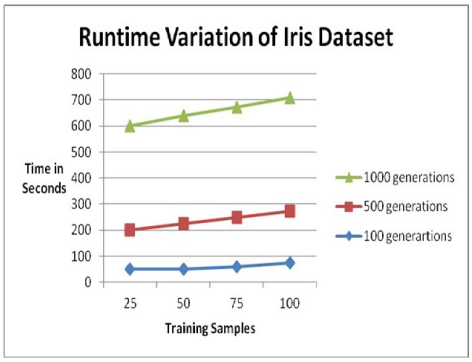

FIGURE 4. Runtime Variation of Iris Dataset

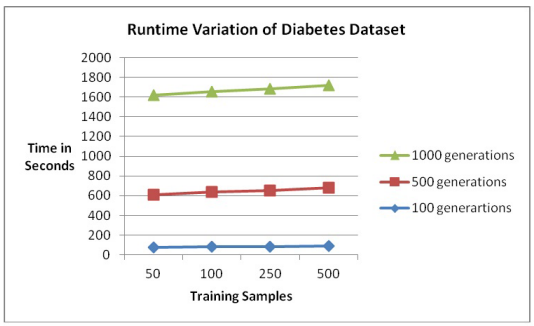

FIGURE 5. Runtime Variation of Diabetes Dataset

CONCLUSION

In this paper, a privacy preserving ant colony optimization based neural learning classifier is proposed. The datasets are vertically partitioned into two sets and given to two users, each user has their own set of data. These two users jointly build a neural network which is securely trained using the ant colony optimization based backpropagation algorithm. The learned weights of the neural network are used for classification. There are many interesting aspects

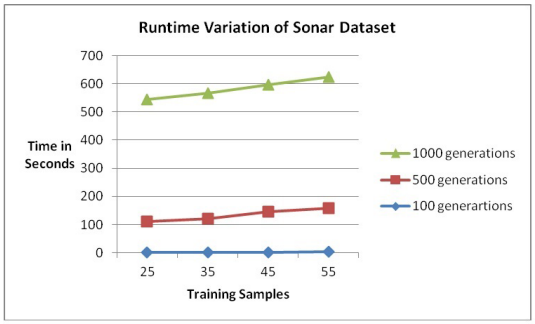

FIGURE 6. Runtime Variation of Sonar Dataset

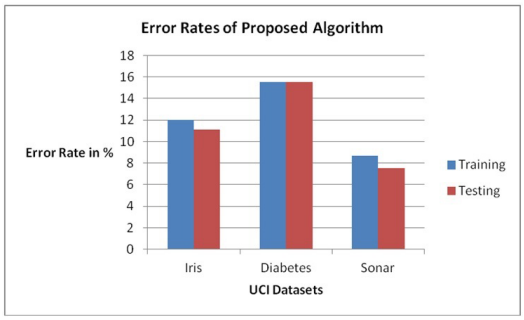

FIGURE 7. Error Rates of Proposed Algorithm

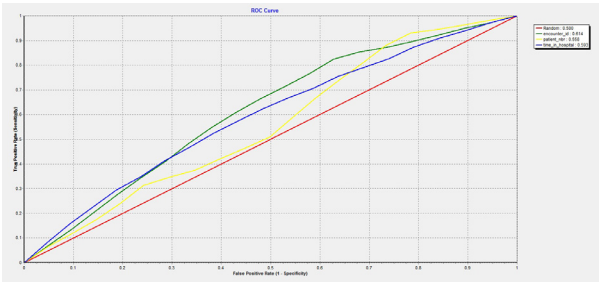

FIGURE 8. ROC Analysis

for future work. The weight optimization can be performed using other metaheuristics algorithm for different classification algorithms.

REFERENCES

[1] G. Nayak and S. Devi, "A survey on privacy preserving data mining: approaches and techniques," *International Journal of Engineering Science and Technology*, vol. 3, no. 3, 2011.

[2] Y. Lindell and B. Pinkas, "Privacy preserving data mining." *Journal of cryptology*, vol. 15, no. 3, 2002.

[3] M. S. Sayyad and P. Kulkarni, "Privacy preserving back propagation algorithm for distributed neural network learning," *International Journal for scientific and Research Publication*, vol. 2, no. 3, pp. 133–6, 2012.

[4] Y. Zhang and S. Zhong, "A privacy-preserving algorithm for distributed training of neural network ensembles," *Neural Computing and Applications*, vol. 22, no. 1, pp. 269–282, 2013.

[5] M. Mavrovouniotis and S. Yang, "Training neural networks with ant colony optimization algorithms for pattern classification," *Soft Computing*, vol. 19, no. 6, pp. 1511–1522, 2015.

[6] B. Guijarro-Berdiñas, S. Fernandez-Lorenzo, N. Sánchez-Maroño, and O. Fontenla-Romero, "A privacy-preserving distributed and incremental learning method for intrusion detection," in *International Conference on Artificial Neural Networks*. Springer, 2010, pp. 415–421.

[7] F. Emekçi, O. D. Sahin, D. Agrawal, and A. El Abbadi, "Privacy preserving decision tree learning over multiple parties," *Data & Knowledge Engineering*, vol. 63, no. 2, pp. 348–361, 2007.

[8] P. K. Fong and J. H. Weber-Jahnke, "Privacy preserving decision tree learning using unrealized data sets," *IEEE Transactions on knowledge and Data Engineering*, vol. 24, no. 2, pp. 353–364, 2012.

[9] A. Gangrade and R. Patel, "Privacy preserving two-layer decision tree classifier for multiparty databases," *International Journal of Computer and Information Technology (2277–0764)*, vol. 1, no. 1, pp. 77–82, 2012.

[10] W. Du and Z. Zhan, "Building decision tree classifier on private data," in *Proceedings of the IEEE international conference on Privacy, security and data mining-Volume 14*. Australian Computer Society, Inc., 2002, pp. 1–8.

[11] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[12] J. Vaidya and C. Clifton, "Privacy-preserving decision trees over vertically partitioned data," in *IFIP Annual Conference on Data and Applications Security and Privacy*. Springer, 2005, pp. 139–152.

[13] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612–613, 1979.

[14] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Y. Zhu, "Tools for privacy preserving distributed data mining," *ACM Sigkdd Explorations Newsletter*, vol. 4, no. 2, pp. 28–34, 2002.

[15] W. Fang and B. Yang, "Privacy preserving decision tree learning over vertically partitioned data," in *Computer Science and Software Engineering, 2008 International Conference on*, vol. 3. IEEE, 2008, pp. 1049–1052.

[16] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, "State-of-the-art in privacy preserving data mining," *ACM Sigmod Record*, vol. 33, no. 1, pp. 50–57, 2004.

[17] C. C. Aggarwal and S. Y. Philip, "A condensation

approach to privacy preserving data mining," in *International Conference on Extending Database Technology*. Springer, 2004, pp. 183–199.

[18] B. Pinkas, "Cryptographic techniques for privacy-preserving data mining," *ACM Sigkdd Explorations Newsletter*, vol. 4, no. 2, pp. 12–19, 2002.

[19] Y. Li, M. Chen, Q. Li, and W. Zhang, "Enabling multilevel trust in privacy preserving data mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 9, pp. 1598–1612, 2012.

[20] K.-P. Lin and M.-S. Chen, "On the design and analysis of the privacy-preserving svm classifier," *IEEE Transactions on Knowledge and Data Engineering*, vol. 23, no. 11, pp. 1704–1717, 2011.

[21] B. Guijarro-Berdiñas, D. Martínez-Rego, and S. Fernández-Lorenzo, "Privacy-preserving distributed learning based on genetic algorithms and artificial neural networks," in *International Work-Conference on Artificial Neural Networks*. Springer, 2009, pp. 195–202.

[22] T. ElGamal, "A public key cryptosystem and a signature scheme based on discrete logarithms," *IEEE transactions on information theory*, vol. 31, no. 4, pp. 469–472, 1985.

[23] S. Rajasekaran and G. V. Pai, *Neural networks, fuzzy logic and genetic algorithm: synthesis and applications (with cd)*. PHI Learning Pvt. Ltd., 2003.