2011 Third International Conference on Computational Intelligence, Modelling & Simulation

An Improved Two-Step Supervised Learning Artificial Neural Network for Imbalanced Dataset Problems

Hasrul Che Shamsudin, Asrul Adam, Mohd Ibrahim Shapiai,
Mohd Ariffanan Mohd Basri, Zuwairie Ibrahim and Marzuki Khalid

Faculty of Electrical Engineering
Universiti Teknologi Malaysia
81300 Skudai, Johor, Malaysia
Centre of Artificial Intelligent and Robotics (CAIRO),
Universiti Teknologi Malaysia,
Jalan Semarak, 54100, Kuala Lumpur, Malaysia
E-mail: hasrulcheshamsudin@yahoo.com, asruldm@fkegraduate.utm.my, ibrahim@fke.utm.my,
ariffanan@fke.utm.my, zuwairie@fke.utm.my, marzuki@utm.my

***Abstract*— An improved two-step supervised learning algorithm of Artificial Neural Networks (ANN) for imbalanced dataset problems is proposed in this paper. Particle swarm optimization (PSO) is utilized as ANN learning mechanism for first step and second step. The fitness function for both steps is Geometric Mean (G-Mean). Firstly, the best weights on network are determined with a decision threshold is set to 0.5. After the first step learning is accomplished, the best weights will be used for second step learning. The best weights with the best value of decision threshold are obtained and can be used to predict an imbalanced dataset. Haberman's Survival datasets, which is available in UCI Machine Learning Repository, is chosen as a case study. G-Mean is chosen as the evaluation method to define the classifier's performance for a case study. Consequently, the proposed approach is able to overcome imbalanced dataset problems with better G-Mean value compared to the previously proposed ANN.**

Keywords- artificial neural network; imbalanced dataset problems; particle swarm optimization; binary classification

I. INTRODUCTION

ANN is an information processing paradigm that is inspired by a biological human nervous system. Many applications have been developed using ANN algorithm and most of the applications are on predicting future events based on historical data. The efficient applications of ANN depend on the learning algorithm. Back Propagation (BP) ANN based on gradient descent where weights change once in each cycle after the entire input sample have been trained are commonly used for classification problem.

However, the major disadvantage of BP are it converge rate relatively slow [1] and being trapped at local minima especially for those non-linearly separable pattern classification problems or complex function approximation problem [2].

Currently, PSO has been used to train ANN. PSO is robust stochastic optimization technique that is introduced by Eberhart and Kennedy in 1995. In 1998, Shi and Eberhart firstly introduced an inertia weight, *w* into the previous PSO algorithm [3]. The performances of PSO algorithm can be improved by adjusting inertial weight which is known as Adaptive Particle Swarm Optimization algorithm (APSO). The inertial weight is adaptively changed in different searching stages for the improved algorithm. On the other hand, a study carried out by Zhang *et al.* [4], showed that a hybrid PSO-BP performs better than conventional ANN in convergent speed and convergent accuracy.

An imbalanced dataset can be defined as a complete data set that consisting unsymmetrical distribution, that is to say, a minority class which is significantly less than a majority class. Conventional ANN unable to learn with wrongly responds to minority class of imbalanced dataset. This problem happens because the overwhelming samples in majority classes affect the adjustment procedure during training [5].

Nowadays, learning from imbalanced datasets has difficulty to be solved. Moreover, imbalanced datasets exist in many real applications such as computer security [6], biomedical [7-8], remote-sensing [9], engineering [5-6], and manufacturing industries [7]. Meanwhile, many learning algorithms have been proposed to improve the performance of conventional classifiers, such as Decision Tree (C4.5) [8], Bayesian Network (BN) [9], ANN [10-14], Mimimax Probability Machine (MPM) [15], and Support Vector Machine (SVM) [16]. Generally, two different approaches are commonly used by researchers for handling the problems which are data level and algorithmic level [17]. At the data level, over sampling and under sampling method modify the prior probability to minimize the imbalanced effect [18]. On the other hand, the algorithmic level involves some modifications on internal algorithm.

This paper utilizes ANN as a tool for classification of imbalanced datasets. This is due to the effectiveness of ANN in modeling real world complex problems [8-9, 19-24].

978-0-7695-4562-2/11 $26.00 © 2011 IEEE
DOI 10.1109/CIMSim.2011.28

In order to improve the performance of ANN for imbalanced datasets, this paper enhances the previous existing approach [13] which namely a modified ANN learning algorithm. Previously, a conventional Levenberg-Marquadt (LM) backpropagation learning algorithm is used in first step learning where mean square error (MSE) acts as fitness function. While in second step learning, PSO with G-Mean function acts to find the best value of decision threshold. This paper enhances the previous approach with utilizes PSO as learning mechanism in first step learning. The fitness function for first step learning is changed to G-Mean. The learning mechanism for second step learning is maintained based on previous approach.

The rest of this paper is organized as follows. Section 2 focuses on a conventional ANN and evaluation method the problems. Section 3 describes the proposed two-step learning approach based on ANN for the problems. In Section 4, several experimental results are presented and discussed. Finally, conclusions are given in Section 5.

II. Conventional ANN and Performance Measure for Imbalanced Data Set Problems

A. ANN for Binary Classification Problem

A single layer feedforward ANN consists of a set of interconnected processing units known as node, neurons or cells as shown in Fig. 1. The weights, *w*, are located on network between all links from input layer to hidden layer and from hidden layer to output layer. The output from the network is compared with actual to get the desired output from ANN. In order to match the output, network will update the connection weights to get the best output. BP is a common learning technique that is used to minimize Mean Square Error (MSE) between the actual outputs of the network and the desired outputs. Each node has activation functions and the common activation functions are linear and sigmoid function. The activation signal sent (output) by each node to others node travel through weighted connection and each of these nodes accumulates the received inputs, producing an output according to an internal activation function. A sigmoid function, $f(x)$, shown in (1) used at the output layer to calculate the limit value of the desired output between 0 and 1.

$$f(x) = \frac{1}{1+e^{-x}} \tag{1}$$

For binary classification problem, a step function, $g(f(x))$, shown in (2), is used to group the class, which is either 0 or 1, based on the threshold value, Θ, as the decision threshold.

$$g(f(x)) = \begin{cases} 1 & if \quad f(x) > \theta \\ 0 & otherwise \end{cases} \tag{2}$$

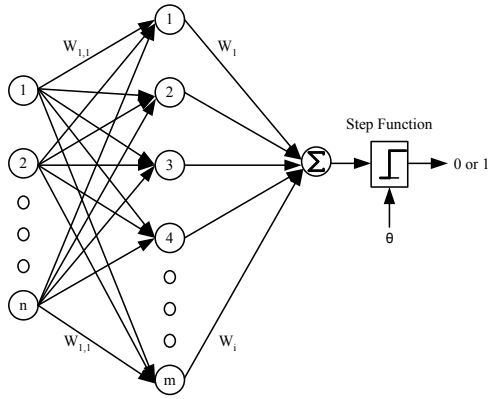

Figure 1. Architecture of feedforward ANN for binary classification.

B. Evaluation Method

G-Mean is one of standard evaluation methods used to measure the performance of an imbalanced dataset classifier [25]. The reason of using G-Mean is to balance the ratio of prediction between majority and minority class. The percentage of G-Mean indicates that how good an imbalanced dataset classifier to predict majority and minority class. The G-Mean is calculated as follows:

$$G-Mean = \sqrt{(TNR \times TPR)} \tag{3}$$

where,

$$TNR = \frac{TN}{TN+FP} \tag{4}$$

$$TPR = \frac{TP}{TP+FN} \tag{5}$$

The *TP*, *FN*, *FP*, and *FN*, can be defined as follows. True Positive (*TP*) refers to correctly prediction of the majority class. False Negative (*FN*) refers to wrongly prediction of the minority class as majority class. False Positive (*FP*) refers to wrongly prediction of majority class as minority class. True Negative (*TN*) refers to correctly prediction of minority class.

III. The Improved Two-Step Supervised Learning of Artificial Neural Network

A. Architecture

The proposed approach can be divided into three phases; first-step learning, second-step learning, and testing. Firstly, the required dataset to represent the problem is divided for training and testing process. The class distribution in the complete data consist training and testing data set. The overview of the previously proposed approach [13] is shown in Fig. 2 while an improved two-step supervised learning

ANN is shown in Fig. 3. The different between architecture in Fig. 2 and Fig. 3 is the learning mechanism in first step learning. As mentioned in Section I, architecture in Fig. 2 utilizes LM backpropagation learning with MSE while architecture in Fig. 3 utilizes PSO algorithm with G-Mean to find the best weights on network.

B. *First Step Learning*

This section describes the first step learning mechanism of the proposed ANN classifier. Single hidden layer with 10 neurons ANN topology is chosen for the first step learning. APSO algorithm is used for training the network [26] while G-Mean as fitness function for the APSO. The idea of the APSO is to search the global best solution for the network weights by maintaining the output threshold, θ to 0.5. The trained network consists all the best weights, w_i, that are used as the input to the second step learning and testing process.

C. *Second Step Learning*

The main purpose of the second step learning mechanism is to further optimize the trained network. With the best weights gained from the first step learning, APSO is employed again to find the optimum threshold value, $\theta_{optimized}$ during the second step learning, as shown in Fig. 4. Note that, inertial weight for the APSO is changed adaptively in different searching stages had been used for both learning step. Similarly, G-Mean is used as the fitness function to tune the decision threshold, θ. PSO setting parameters used for both learning phase is shown in Table I.

D. *Testing*

As shown in Fig. 2, the output of testing data set (unseen data) is predicted using the ANN after the first step and the second step learning are accomplished. Finally, G-Mean is calculated to evaluate the classifier's performance for imbalanced data set.

IV. Experimental Results and Discussions.

A. *Haberman's Survival Dataset*

Table II shows a Haberman's survival data set that is taken from UCI Machine Learning Repository [12]. This data set consists of three attributes for input and one output. *Attribute 1* values range from 30 to 83, *Attribute 2* values range from 58 to 69, and *Attribute 3* values range from 0 to 52. The output is in categorical form, which is either 0 or 1. There are 306 collected samples in this data set. 74.5 percent of samples are 0 (majority class) whereas 26.5 percent of samples are 1 (minority class).

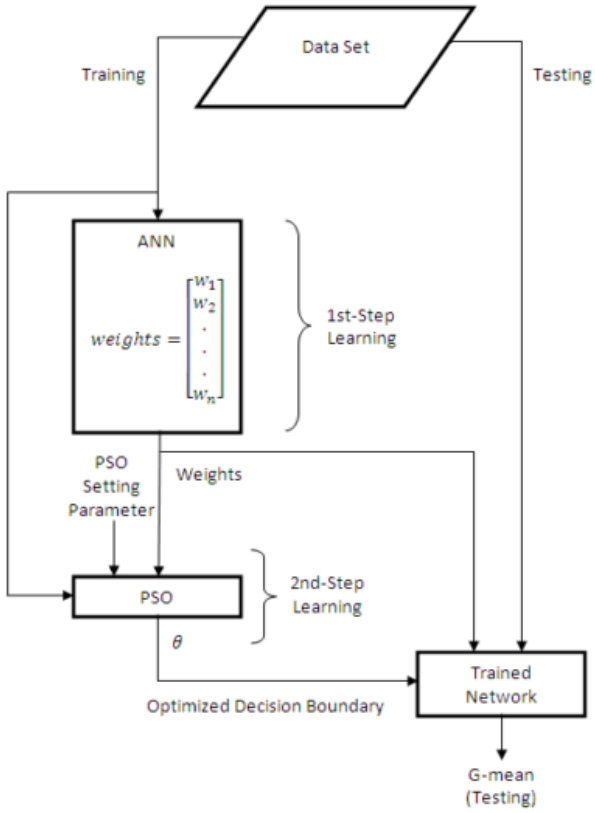

Figure 2. Architecture of the two-step learning ANN model for imbalanced dataset problems [13].

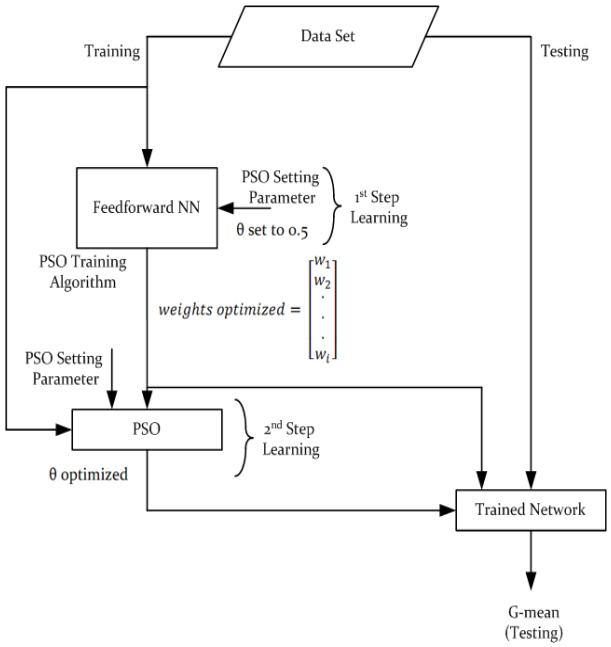

Figure 3. Architecture of the improved two-step supervised learning of ANN for imbalanced dataset problems.

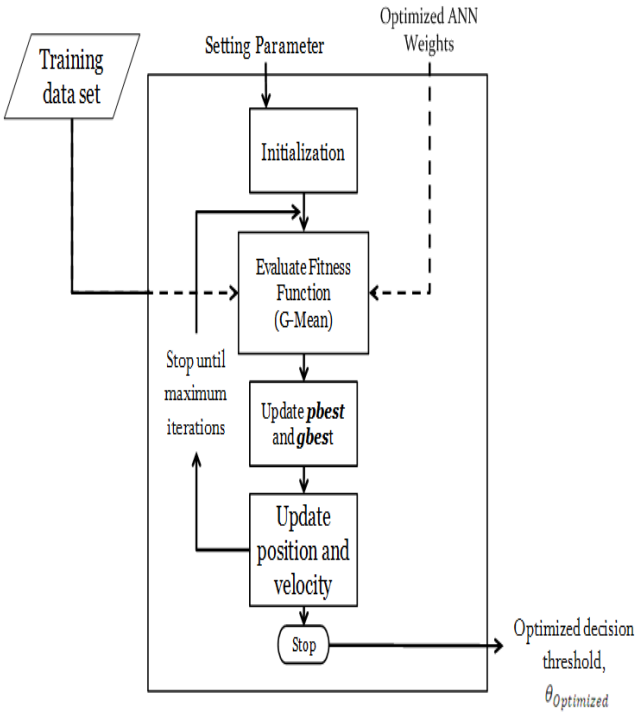

Figure 4. Implementation of PSO algorithm during the second-step learning mechanism of the proposed ANN.

TABLE I. PSO setting parameter

Number of particles	10
Decrease inertia weight, ω	0.9~0.4
Cognitive coefficient, C_1	2
Social coefficient, C_2	2
r_1 and r_2	Random [0,1]
Maximum iteration	100 for each step

TABLE II. Haberman's survival diabetes dataset

Unit ID	Attributes			Output
	1	2	3	
1	30	64	1	0
2	30	62	3	0
3	30	65	0	0
…..	…..	…..	…..	…..
306	83	58	2	1

B. Implementation of the Proposed Approach to Haberman's Survival Dataset

A high level implementation of the proposed approach is shown in Fig. 5. The classifier has three inputs, which are the age of patient at time of operation, the patient's year of operation (year-1990), and the number of positive auxiliary nodes detected. The output represents the status of patient, whether the patient died or survived. According to this dataset, 73.5% of patients were survived (majority class) and the remaining were not survived (minority class). The class distribution can be written in a mathematical term, majority: minority = 0.735: 0.265 to represent the tested patient. In order to compare the performance with the conventional ANN, an improved two-step supervised learning ANN was executed 50 times and the result are shown in Table III.

There are G-Mean train and G-Mean Test values shown in Table III to indicate the performance after training and testing for each learning algorithm. According to Table III, the proposed approach performance about double G-Mean test value than the conventional ANN. In addition, the proposed technique provides better prediction in term to make the training process more reliable, as it provided a more stable behavior as indicated by smaller value of standard deviation. Besides that, Table III shows the proposed approach performance better than previous two-step learning mechanism ANN [13].

V. CONCLUSION

In this study, PSO was successfully applied in neural network and has been tested using Haberman's survival dataset. The ANN learning algorithm separately learns that namely two-step learning mechanism, which is consists of the parameter tuning of ANN's weights and optimizing the decision threshold.

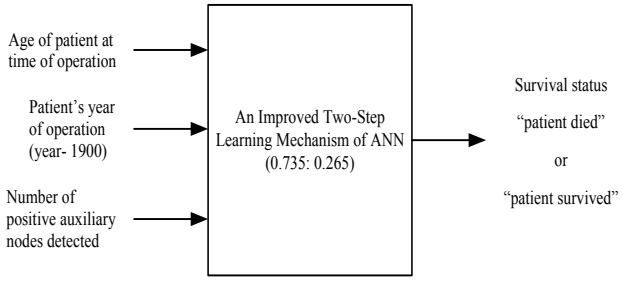

Figure 5. High level implementation of the proposed approach to Pima Indians Diabetes data set.

Most important finding in this study is the improved two-step supervised learning ANN algorithm provides better classification performance compared to the conventional ANN and the previously proposed ANN [13]. For future works, the two-step supervised learning ANN algorithm can be enhanced by using a hybrid PSO algorithm combining the adaptive PSO algorithm with back-propagation algorithm, which is to combine the strong ability in global search and strong ability in local search as suggested by Zhang [26] to solving imbalanced data set problem.

ACKNOWLEDGEMENT

This work was financially supported by the UTM-INTEL Research Collaboration (Vote 73332) and the Ministry of Higher Education (MOHE) Fundamental Research Grant Scheme (FRGS) (Vote 78645).

Table III: Comparison of the average G-Mean using the conventional ANN, modified ANN, and the proposed two-step learning ANN based on Haberman's survival data set.

Classifier	Conventional ANN, $\theta = 0.5$ [13]		Modified ANN [13]		The proposed ANN	
Measurement	G-Mean Train (%)	G-Mean Test (%)	G-Mean Train (%)	G-Mean Test (%)	G-Mean Train (%)	G-Mean Test (%)
Average	38.87	36.04	71.26	58.67	71.73	65.23
Maximum Score	64.63	59.16	80.16	70.47	76.56	76
Standard Deviation	12.78	11.35	4.39	4.91	2.41	4.01

REFERENCES

[1] Zweiri T.H., Whidborne J. F., Sceviratne L.D., "A Three-Term Backpropagation Algorithm", Neurocomputing, Vol. 50, pp. 305-318.

[2] Marco Gori, Alberto Tesi, "On the Problem of Local Minima in Back-Propagation", IEEE Transactions on Pattern Anal. Mach. Intell, Vol. 14, pp. 76-86, 1992.

[3] Y. Shi, R. C. Eberhart, "A Modified Particle Swarm Optimizer", Proceeding of IEEE World Conference on Computation Intelligence, pp. 69-73, 1998.

[4] Jing-Ru Zhang, Jun Zhang, Tat-Ming Lok, Micheal R. Lyu, "A Hybrid Particle Swarm Optimization-Back-Propagation Algorithm for Feeddforward Neural Network Training", Applied Mathematics and Computation, pp. 1026-1037, 2007.

[5] Yi L. Murphey, Haoxing Wang, Guobin Ou, Lee A. Feldkamp, "OAHO: an Effective Algorithm for Multi Class Learning from Imbalanced Data", Proceedings of IEEE, International Joint Conference on Neural Networks, Orlando, Florida, USA, 2007.

[6] D. Cieslak, N. Chawla, and A. Striegel, "Combating Imbalance in Network Intrusion Datasets", Proceedings of IEEE International Conference on Granular Computing, pp. 732-737, 2006.

[7] Maciej A. Mazurowskia, Piotr A. Habasa, Jacek M. Zuradaa, Joseph Y. Lob, Jay A. Bakerb, Georgia D. Tourassib, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance", Advanced in Neural Networks Research: International Joint Conference on Neural Networks (IJCNN '07), Vol. 21, pp. 427-436, 2008.

[8] B. Anuradha and V. C. Veera Reddy, "ANN for Clasification of Cardiac Arrhythmias", Asian Research Publishing Network (ARPN) Journal of Engineering and Applied Sciences, Vol. 3, No. 3, 2008.

[9] L. Bruzzone, S.B. Serpico, "A Classification of imbalanced remote-sensing data by neural networks", Pattern Recognition Letters, Vol. 18, pp. 1323-1328, 1997.

[10] Giang H. Nguyen, Abdesselam Bouzerdoum, and Son L. Phung, "A Supervised Learning Approach for Imbalanced Data Sets", Proceedings of 19th International Conference on Pattern Recognition (ICPR 2008), pp. 1-4, 2008.

[11] Z.Q. Zhao, "A novel modular neural network for imbalanced classification problems", Pattern Recognition Letters, Vol. 30, pp. 783-788, 2008.

[12] Asuncion, A. & Newman, D.J. (2007). UCI Machine Learning Repository [http://www.ics.uci.edu/~ mlearn/MLRepository.html]. Irvine, CA: University of California, School of Information and Computer Science.

[13] Asrul Adam, Ibrahim Shapiai, Zuwairie Ibrahim, Marzuki Khalid, "A Modified Artificial Neural Network Learning Algorithm for Imbalanced Data Set Problem", Proceedings of International Conference on Computational Intelligence, Communication Systems and Networks, Vol. 2, pp. 44-38,2010.

[14] Anand R., Mehrotra K. G., Mohan C. K., Ranka S., "An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets", IEEE Transactions on Neural Networks, Vol.4, No.6, November 1993.

[15] Kaizhu Huang, Haiqin Yang, Irwin King, Michael R. Lyu, "Learning Classifiers from Imbalanced Data Based on Biased Minimax Probability Machine", Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR04), Vol.2, pp. 558-563, 2004.

[16] Yuchun Tang, Yan-Qing Zhang, Nitesh V. Chawla, and Sven Krasser, "SVMs Modeling for Highly Imbalanced Classification", IEEE Transaction on System, Man, and Cybernetics, Vol. 39, pp. 281-288, 2008.

[17] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, "Handling imbalanced datasets: A review",GESTS International Transactions on Computer Science and Engineering, Vol. 30, 2006.

[18] R. Alejo, V. Garcia1, J.M. Sotoca, R.A. Mollineda and J.S. Snchez, "Improving the Classification Accuracy of RBF and MLP Neural Networks Trained with Imbalanced Samples", Lecture Note on Computer Science (LNCS), Intelligent Data Engineering and Automated Learning, Vol. 4224, pp. 464-471, 2006.

[19] Guoqiang Peter Zhang, "Neural Networks for Classification: A Survey", IEEE Transactions on Systems, MAN, and Cybernetics, Vol. 30, No. 4, pp. 451-462, 2000.

[20] B. Yegnanarayana, Artificial neural networks for pattern recognition, Springer India, in co-publication with Indian Academy of Sciences, Vol. 19, No. 2, pp. 189-238, 1994.

[21] Cochocki, A. and Unbehauen, Rolf, "Neural Networks for Optimization and Signal Processing", John Wiley & Sons, Inc., New York, NY, USA, 1993.

[22] Wenjun Zhang and Albert Barrion, "Function Approximation and Documentation of Sampling Data Using Artificial Neural Networks", Environmental Monitoring and Assessment, Springer Nether-lands, Vol. 122, No. 1-3, pp. 185-201, 2006.

[23] Rangachari h. , Kishan G. Mehrotra, Chilukuri K. Mohan, and Sanjay Ranka, "An Improved Algorithm for Neural Network Classification of Imbalanced Training Sets", IEEE Transactions on Neural Networks, Vol. 4, No. 6, pp. 962, 1993.

[24] Won Kim, Ick Choy, and Gwi-Tae Park, "Sensorless Speed Control System Using a Neural Network", International Journal of Control, Automation, and Systems, Vol. 3, No. 4, pp. 612-619, 2005.

[25] Cheng G. Weng and Josiah Poon, "A New Evaluation Measure for Imbalanced Datasets", 2006.

[26] Jing-Ru Zhang, Jun Zhang, Tat-Ming Lok, Micheal R. Lyu, "A Hybrid Particle Swarm Optimization-Back-Propagation Algorithm for Feedforward Neural Network Training", Applied Mathematics and Computation, pp. 1026-1037, 2007.