Investigating Sensitivity of Nonlinear Classifiers by Reducing Mean Square Error (MSE)

[1]Abeer A. Al-Mohdar, [2]Ahmed A. Bahashwan

[1]Assistant Professor, Dept. of Information Technology, Faculty of Computers & Information, Hadhramout University, Yemen
[2]Faculty of Engineering & Petroleum, Dept. of Electronics & Communication Engineering, Hadhramout University, Yemen

Abstract - **In this paper a new method is presented for handling the problems of Artificial Neural Networks (ANNs) in a Self-Organizing Map (SOM), the problems are the presence of noise and missing in datasets. Based on a robust and speedup of the investigation of a nonlinear classifications in SOM and Adaptive Linear Neuron ANNs (Adaline) algorithms that dealing with missing or noise data a new model proposed. the main objective of the proposed model is to contribute the advantages of all methods (SOM with Adaline) in rising the quality of SOM to handle the noise and missing value to improve its sensitivity of our model using combing the advantages of all methods (SOM with Adaline) in anew algorithm abbreviation (ADA-SOM). Simulation results show that the algorithm ADA-SOM achieved better performance and higher sensitivity to accurate MSE (Mean Square Error) than other standard classifiers, our model gets accurate benchmark results.**

Keywords: Self-organizing-map, Adaptive Linear Neuron, missing data, Artificial Neural Networks, Unsupervised classifiers.

I. INTRODUCTION

ANNs are accurate electronic models to solve problems of clustering and classification algorithms through experience and learning from large data sets. Simulating a functioning of a biological brain, which have the ability to store knowledge and self-learning in both sides (Supervised or Unsupervised learning). The development of the ANN based on updating connection weights between neurons through learning process by reducing the misfit between actual and expected output.

ANNs are also characterized by accuracy and efficiency that depends on the perfection and validity of the databases that used to learn and test the system for any defect in these data (missing or noise) leads to a defect in building the system. One of the most important algorithms in ANNs is SOM [1- 4], its characterized by ability to deal with large databases with the presence of obstacles that affect its performance in the correct clustering of data due to the

presence of missing or noise data that leads to an incorrect representation of data and thus obtaining inaccurate system.

To solve this problem, we proposed a new technique based on the properties of Adaptive Linear Neuron (Adaline) algorithm, [5, 6] which is characterized by ability to deal with missing and noise of data [30, 31, 19]. This technique operates in a small scale and based on statistical methods that have enhanced its performance.

Adaline was considered one of the best algorithms that deal with missing values by giving the minimum error rate (MER), this issue is a motivation to enhance the work of the SOM algorithm that has high efficiency and accuracy in performance required in validation and completeness of data. Also calculating the sensitivity of the network's output enhances the effectiveness of results. The main aim of this paper is to contribute in the design of an adaptive architecture (ADA-SOM) based on Adaline sensitivity by retraining our model.

Adaptive Linear Neuron (Adaline) is a simple linear ANN model that has two layers, every neuron in each layers interacts to each other and transfer the information between them where knowledge is not stored in the individual nodes, but rather it's represented by the weights of the connections between the nodes. Adaline network exercise the result with Least Mean Squares (LMS) algorithm for learning. The scope of Adaline network is to recognize patterns, data filtering, or to approximate linear function [7]. Missing and noise data are the main objectives in the most researches as well as the detection of ambiguity elements when analyzing a huge of data.

In this paper a new method is introduced that contributes to solve the problems of missing and noise in a data set. The method uses the technique and properties of SOM to take advantage of its benefits in dealing with missing imputation data sets combing with the special Adaline technique as a new method ADA-SOM. We need to solve them appropriately in order to provide an efficient result. Most scientific research try to solve the problem of missing values and noise data domains

such as data mining [8,9], medicine learning [10-12], statistics [13-15], and bioinformatics [16-18].

The organization of the rest of this paper is as follows. Section 3 describes the principle of the SOM techniques, including the Adaline algorithm. Section 4 describes the proposed algorithm in a new structure ADA-SOM. Section 5 describes the experimental Results .Finally, the conclusion will be found in section 6.

The Self-Organizing Map (SOM) algorithm is based on an unsupervised learning principle, where training the entirely data-driven and no information about the output data is required. SOM is a collection of n reference vectors organized in a neighborhood network, and they have the same dimension as the input vectors. Neighborhood function is usually given in terms of a two-dimensional neighborhood matrix W(i,j). In a two-dimensional map, each node has the same neighborhood radius, which decreases linearly to zero during the SOM learning. The conventional Euclidian distance is used to determine the Best-Matching Unit (BMU) (so called 'winner') on a map for the input vector Vmn.

The goal is to discover some underlying structure of the data. Kohonen's SOM is also called a topology-preserving map because there is a topological structure imposed on the nodes in the network. Each node in a given layer has been identical in that each is connected with all of the nodes in the upper and/or lower layer. In the brain, neurons tend to cluster in groups. The connections within the group are much greater than the connections with the neurons outside of the group. Kohonen's network tries to mimic this in a simple way[19].

1.1 The SOM algorithm can be summarized as follows

Neurons in input layer V_{mn} has weights as the length T weight matrix $W_i = [W_{i\,1}, W_{i\,2}, ..., W_{i\,3}]$, of the learning data samples, V_{mn}, for n = [1, 2, ..., N]. The output layer is connected in an array (usually 1 or 2 dimensional) fully connected (i.e. all nodes in the input layer are connected to all nodes in the output layer). V_{mn} shaped as a square lattice. The network name as competitive learning algorithm as follows steps:

Step 1: initialization network as randomly choosing an input vector V_{mn} .

Step 2: Computes the "BMU output node i, of minimum Distance usual measure as Euclidean Distance (ED) in "Equation (1)":

$$ED = \min_j \|V - W_j(p)\| = \sqrt{\sum_{i=1}^{n} [V_i - W_{ij}(p)]^2} \dots (1)$$

where $j = 1, 2, \cdots, m$

Step 3: Update the winner unit BMU as shown in "Equation (2)":

$$W_{ij}(p + 1) = \Delta W(p) + \alpha(P)\left(V(p) - W_{ij}(p)\right) \dots (2)$$

$$\Delta W_{ij}(p) = \begin{cases} \alpha\,[V_i - W_{ij}(p)], & j \in \Lambda_j(p) \\ 0, & j \notin \Lambda_j(p) \end{cases}$$

Where α is the learning rate and Λj (p) is the neighborhood function centered in Vi at iteration p.

Step 4: Repeat step 2 until changing all feature map. After processing all of the input, the result should be a spatial organization of the input data organized into similar regions.

1.2 The Adaline algorithm

Adaline discovered by Widrow & Hoff (1960), is an ANN that has a weight updating algorithm $\Delta W_{ij}(p)$ and an activation function has multiple inputs, architecture and biases that are fully connected directly to an output neuron its behavior represented in Fig-1 below. During the training of the Adaline, we have used the identity of activation function. Learning rate (α) parameters is used in training Adaline algorithm.

The main objective of the α is to fix the weights that connect the input V_i and output units so that the difference between the input for the output unit (Y) and the target value (t) becomes minimal. Adaline can also be combined so that the output from some Adaline becomes input to other Adaline, and will form a multilayer network called the Madaline (Many Adaptive Linear Neurons)[1]. The learning process of the algorithm depends on three steps:

Step 1: Dataset are input to the network, and then can perform the calculation of weight. Query calculation is a process whereby the algorithm works to generate new weights that are used as a rule or reference to determine the classes in new output data must give small MSE (Mean Square Error).

Step 2: After the new weights have been successfully saved, the algorithm can enter/add a new input data and see the result of the class determination automatically based on the calculation of the test value and the written test score and the new weight generated in the previous stage stands on the value of MSE to get pure data set.

MSE is a measure of the differences between the predicate output values and actual output values for the dataset being considered. It provides a means of quantifying these

differences. For a given dataset, MSE can be computed as follows:

$$\text{MSE} = \frac{1}{n}\sum_{-}(v_i - \bar{v}_i)^2 \dots (3)$$

Step 3: Replace a variable of a missing in the prediction steps at best adds a relevant variable to the data, decreasing variance. Because the value can be replaced with a highly accurate estimate, its inclusion in the model is likely to reduce the effectiveness at capturing predictive patterns involving the other variables, as above-mentioned , it must build an accurate base classifier and accurate imputation models for any possible missing values.
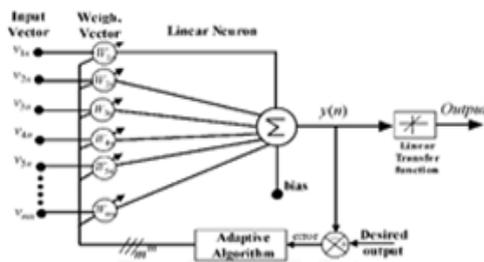

Figure 1: Adaline behavior

Adaline Sensitivity definition: Sensitivity is defined as the probability of the Adaline's output inversion due to the variation on weight for all input points, which can be expressed as $\mathbf{Y_i} = \frac{\mathbf{v_i}}{\bar{\mathbf{v}}_i} \dots (4)$

Where missed word is the number of all input V_i points that cause the Adaline's output changed due to the variation on weight. The sensitivity computation for Adaline based on the variation on input can converted into the variation on weight. For any missed word of the input, and there is the following equivalent transformation relation:

$$V_i\, W = \bar{v}_i\, \Delta W \dots (5)$$

The study [20, 21] shows that the sensitivity of Adaline can be approximately computed as the following:

$$\Delta W = \frac{\frac{\Delta W}{W}}{\propto} = |w'| \dots (6)$$

1.3 Why Classification of missing values it Matters

Missing and noise data considered as a big problematic in a real life, because all science needs integrated database to achieve actual systems. Ignoring missing data can have an impact on the results of the system analysis [22, 23].

Discarded the missing data has advantages in a big database that make it easy to use, and very simple. It is the

default in most statistical packages but it has limitations. It can substantially lower the sample size, leading to a severe lack of power. This is especially true if there are many variables involved in the analysis, each with data missing for a few cases. It can also lead to biased and wrong results, depending on why the data are missing and an impact on the results of the system analysis [10].

II. THE PROPOSED METHOD (ADA-SOM)

In the proposed algorithm, we relied on two basic stages:

1] Pre-processing (clustering): In this stage, data was initialized and processed from lost, and noise data depending on the performance of our proposed Adeline algorithm:

Training Adaline classifier based on updating W_i^T recursively for each input vector V_i.

$$Y_i = W_i^T V_i \dots (7)$$

Applying LMS steepest-descent-based to get the minimum of MSE:

$$J_{MSE}(w_i) = \sum_{i=1}^{N} E\{e_i^2\} = \sum_{i=1}^{N} E\{Y_{i-} W_i^T V_i^2\} Yi \dots (8)$$

E: expectation operator, e_i: error of each iteration.

Updating W_i^T recursively in negative direction in this gradient:

$$W_{i+1}^T = w_i + \alpha e_i V_i \dots (9)$$

Where α is learning rate which controls the convergence rate that raise up the efficiency of our method.

Repeat step b to refine all input data V_i which stands on the value of MSE to get pure data set .

Compute the Mod equation of the missing data and mutation V_i to $\bar{v}_i$ with mine_i^2.

2] Post-processing (classification): in this stage the dataset ($\bar{\mathbf{v}}_i$) is ready to deal with and benefit from the capacity of our proposed ADA-SOM algorithm, its speed, and its accuracy in classification. Input $\bar{v}_i$ with MSE of min ED in Equation(1). e_i^2 to

Update the BMU of a winner unit to of SOM lattice algorithm as shown in Equation (2) above.

Repeat step 2 until changing all feature map.

After processing all of the input, the result should be a spatial organization of the input data organized into similar regions of all dataset $\bar{v}_i$ gave best classifier with high correlation based

on small MSE of each output. with high sensitivity value in adaption architecture of w′, where is the number of all input V_i points, and is the number of the input points that cause the Adaline's output changed due to the variation on weight in Equation. (6).
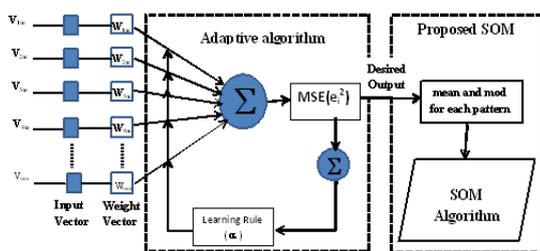

Figure 2: Proposed method (ADA-SOM)

III. EXPERIMENTS

In order to treat errors resulting from missing and noise data, we evaluated the performance of our model in classification algorithm by using two datasets:

1] Hepatitis C Virus (HCV) for Egyptian patients Data Set Number of Instances: 1385(big dataset), Number of attribute: 29 with 7 classes as shown below in Table1 [23].

Table 1: HCV Dataset Description

Attribute	type	Attribute	type	Attribute	type
Age	int	Jaundic	int	ALT 12	int
Gender	int	Epigast	int	ALT 24	float
BMI	int	pain	int	ALT 36	int
Fever	int	WBC	int	ALT 48	int
Nausea /Vomting	int	RBC	float	ALT after 24	int
Headache	int	HGB	int	Base RNA 4	int
Diarrhea	int	Plat	flt	RNA 12	int
Fatigue &general	int	AST	int	RNA EOT	int
bone	int	ALT 1	int	RNA EF	int
ache	int	ALT4	int		

2] The Monk's problem was the basis of a first international comparison of learning algorithms. One of the Monk's problems has noise added.

Table 2: Momk-1 Dataset Description

Attribute	Type
A1	integer
A2	integer
A3	integer
A4	integer
A5	integer
A6	Integer

The numbers of instances are 432 (small dataset), the numbers of attributes are: 8 with 2 classes [24] as shown below in Table2.

We used 75% of the dataset for the training phase and 25% for the test phase of our proposed model. For the purpose of experimentation the domain has been splitting into a train and test set. Python language and Weka open source machine learning software was used to achieve this experiment. The experimentations has conducted in a machine with core i5 processor and 8GB RAM.

The experimental results were implemented in small monk-1 and large HVC dataset with 20% noise and missing data. The results obtained after 1000 training and testing cycles was set for all experiments are summarized below in table 3.

A validation sets (MSE, sensitivity)were selected for the classifiers to analyze the performance for each classifiers in our experiments. Using network's architecture ADA-SOM which enhanced by reducing variation on inputs data to adjustments weight as well as was improving the classifier performance that good overall results have been achieved.

Our proposed model ADA-SOM classifier network has gave a minimum error MSE than different standard classifiers algorithm C4.5, ml-perceptron, SOM.

Table 3: The Experimental Results of our Model

ML-Algorithm	Sensitivity		MSE	
	HCV	Monk1	HCV	Monk1
C4.5	85.07	92.85	0.150	0.125
ml-perceptron	90.27	93.120	0.0308	0.1094
SOM	91.168	96.650	0.0938	0.0459
ADA-SOM	94.090	97.091	0.0631	0.0102

In the evaluation of smallest error achievement with the ADA-SOM method the method produced the local approximations results that shown in Fig-3 and reduced errors of our model from predicated estimations and enable best accreted of interpolation the model performances.

Establishing the compensation mechanism of ADA-SOM to improve the performance of ANN (missing and noise data) caused by a new architecture, will enhances the effectiveness and replaces the missing data by mean value of implementing dataset (VIC, Monk) which is more effectiveness.
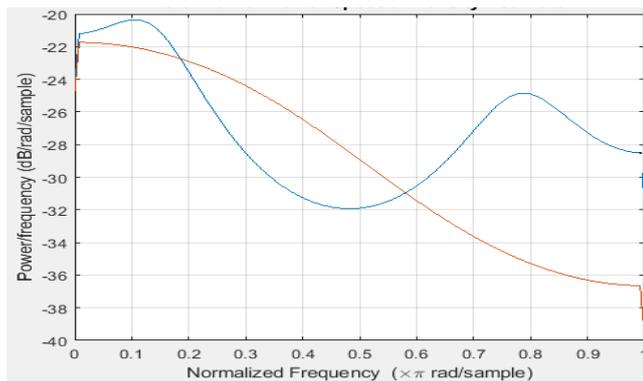
Figure 3: ADA-SOM Local Error Approximations

Based on sensitivity in results of our model, the ADA-SOM was more sensitive in presence of outliers than other classifiers. We obtained higher sensitivity in our model ADA-SOM compared with standard classifiers (C4.5, ml-perceptron, and SOM) as show in Fig-4.
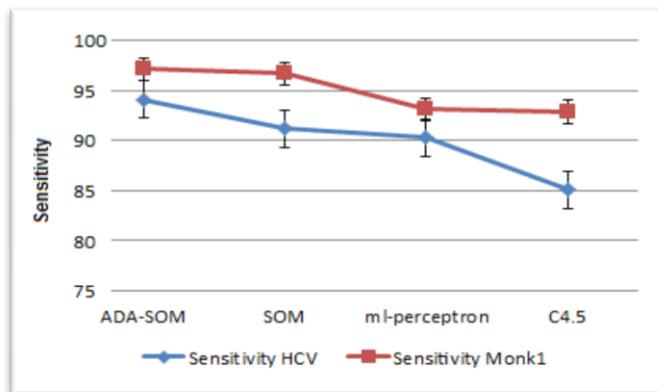

Figure 4: Sensitivity for Classification Algorithms

IV. CONCLUSION

In this paper, we proposed a new method in order to treat with errors resulting from missing and noise data. The performance is measured based on the property of Adaline and SOM sensitivity computation. The proposed algorithm found that the target network ADA-SOM has smaller architecture adaptively. In addition, it reduced the computation of the network implementation effectively. Results also gave lower MSE errors than other standard classifiers algorithm.

The proposed algorithm has controlled the variation on architecture of ADA-SOM caused by creating the sensitivity measure that follows the idea of minimization the error between actual input and predicted output which was more practical in a better performance. Proposed algorithm dataset (VIC, Monk) is more effectiveness. Dealing with literals data will be the next topic to be concerned as a future work to generalize our model.

REFERENCES

[1] T. Kohonen, "Self-Organizing Maps," *Springer-Verlag, Berlin,* 1995.

[2] A. Sorjamaa, B. Maillet, P. Merlin andA. Lendasse,"Some of for finding missing values," *European Symposium on Artificial Neural Networks, Bruges, Belgium,* pp. 25-27, ISBN 2-930307-07-2, 2007.

[3] T.Samad, and S. Harp, "Self-organization with partial data. Network," vol. 3, no. 2, pp. 205–212, 1992.

[4] A. Vellido,K. Gibert, C. Angulo, J.D. Martín Guerrero, "Advances in Self-Organizing Maps, Learning Vector Quantization, Clustering and Data Visualization," *Advances in Intelligent Systems and Computing,* vol. 342, no. 976, DOI:10.1007/978-3-030-19642-4, 2020.

[5] Sun Y, BragaNeto U,ER. Dougherty, "Impact of missing value imputation on classification for dna microarray gene expression data: A model-based study EURASIP," *J Bioinformatics Syst Biol,* vol. 44, no. 1, 2009.

[6] NJ.Horton and KP. Kleinman, "Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models," *The American Statistician* vol. 61, pp. 79-90, 2007.

[7] R. Saranya1, K. Kuppusamy, "Performance analysis of ADALINE and MADALINE Network in classification of Thyroid Disease," *International Journal of Advanced Research Trends in Engineering and Technology (IJARTET),* vol. 3, Special Issue 20, (April) 2016.

[8] JA.Saunders, N. Morrow-Howell,E. Spitznagel, P. Dor, and Proctor EK, et al., "Imputing missing data: A comparison of methods for social work researchers," *Social Work Research,* vol. 30, no. 22, pp. 19-31, 2006.

[9] AWC.Liew,NF. Law,H. Yan, "Missing value imputation for gene expression data: computational techniques to recover missing data from available information," *Briefings in Bioinformatics,* vol. 12, pp. 498-513, 2011.

[10] A. deBrevern,S. Hazout,A. Malpertuy, "Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering," *BMC Bioinformatics,* vol. 5, pp. 1-12, 2004.

[11] D.Wang,Y. Lv,Z. Guo,X. Li,Y. Li, et al., "Effects of replacing the unreliable cdna microarray measurements on the disease classification based on gene expression profiles and functional modules," *Bioinformatics,* vol. 22, pp. 2883-2889, 2006.

[12] HD. Lewis, "Missing data in clinical trials. New England Journal of Lewis HD Missing data in

clinical trials," *New England Journal of Medicine,* vol. 367, pp. 2557-2558, 2012.

[13] Little RJA, Rubin DB, "Statistical Analysis with Missing Data, *Wiley-Interscience,* 2002.

[14] DB. Rubin, "Multiple Imputation for Non-response in Survey," *John Wiley and Sons,* 1987.

[15] JG.Ibrahim, MH. Chen, SR. Lipsitz, AH. Herring,"Missing-data methods for generalized linear models," *Journal of the American Statistical Association,* vol. 100, pp. 332-346, 2005.

[16] T. JOUR, A. Longlong, A. Kenli, A.Canqun, and A. Qi, "A multiple kernel density clustering algorithm for incomplete datasets in bioinformatics," vol. 12, DOI- 10.1186/s12918-018-0630-6, *BMC Systems Biology,* 2018.

[17] C. Ryan, D. Greene andG.Cagney, et al., "Missing value imputation for epistatic MAPs," *BMC Bioinformatics* 11, 197. https://doi.org/10.1186/1471-2105-11-197, 2010.

[18] B. Yang, F. Liu, C. Ren, Z. Ouyang, Z. Xie,X. Bo, and W. Shu, BiRen, "predicting enhancers with a deep-learning-based model using the DNA sequence alone," *Bioinformatics,* vol. 33, no. 13, doi.org/10.1093/bioinformatics/btx105, 2017.

[19] S.Ouadfeul, andL. Aliouane, Lithofacies, "Classification Using the Multilayer Perceptron and the Self-organizing Neural Networks," *Neural Information Processing,* vol. 7667, ISBN: 978-3-642-34499-2, 2012.

[20] J. L. Bernier, J. Ortega and E. Ros, "A Quantitative Study of Fault Tolerance, Noise Immunity, and Generalization Ability of MLPs," *Neural Computation,* vol. 12, no. 12, pp. 2941-2964, 2006.

[21] Ji.Sai, Ping Yang, Shuiming Zhong1, Jin Wang and Jeong-Uk Kim, " A Novel Adaptive Architecture Pruning Algorithm for Madalines*," International Journal of Hybrid Information Technology, International Journal of Hybrid Information Technology* vol. 9, no. 6, pp. 303-316 http://dx.doi.org/10.14257/ijhit.2016.9.6.27.

[22] M. Pampaka, G. Hutcheson, and J. Williams, "Handling missing data: analysis of a challenging data set using multiple imputation," *Int. J. Res. Method Educ.,* vol. 39, no. 1, pp. 19–37, 2016.

[23] https://www.kaggle.com/joelpires/hepatitis-c-virus-for-egyptian-patients-data-set.

[24] https://github.com/nikicc/orange/tree/master/Orange/datasets

****