SOFT MAX ACTIVATION FUNCTION FOR NEURAL NETWORK MULTI CLASS CLASSIFIERS

N.Mohana Sundaram[1], *S.N.Sivanandam*[2]

ABSTRACT

This paper investigates the effectiveness of the Softmax function used as activation function in Neural networks for multi class classification problems. An Elman Neural Network is used with Softmax activation function. The Thyroid Diagnosis problem and Glass type identification problem are considered as multi class datasets which are obtained from UCI machine learning laboratories. The experimental results prove that the Softmax classifiers are more efficient and have higher accuracy than the other methods available in literature.

Key words : Soft-max function, Activation function, Neural Networks, Elman network, Classification, multi class classification.

I. INTRODUCTION

Artificial Neural Networks are efficient tools in constructing the classifiers because of their non linear adaptive learning and dynamic processing capabilities. In this research paper an Elman Neural Network is considered because of its dynamic nature and the non-linear processing capabilities. Classification is a major technique in the data mining which aims at building a classifier to model and identify the major data classes in large datasets. Multiclass classification involves problems with more than two classes to be predicted. Multiclass classification is always more complex than binary classification because several classes may exist, which may lead to increase in the probability of error. Unlike in the case of binary classification problems, it is not just sufficient to choose a threshold value or score to make predictions but the predicted answer is the class (i.e., label) with the highest predicted score. Hence, for multi class classification the Soft max activation function is more suitable.

1.1 Activation function

The activation function also called as the transfer function plays a vital part in neural networks, which transforms or maps the input signals into output signals. Generally in classification applications the output is set at one of two or more levels, depending on whether the total input is greater or less than some threshold value. The activation function of a node defines the output of that node for the given input or set of inputs. Thus the activation function is a decision making function which is commonly a non-linear function.

The commonly used activation functions or transfer functions are

" Linear activation function

" Logistic activation function

[1]Dept. of Computer Science and Engineering, Karpagam Academy of Higher Education, Coimbatore, India.

[2]Professor Emeritus, Karpagam college of Engineering, Coimbatore, India

[1] itismemohan@gmail.com

" Sigmoid or hyperbolic activation function

" Soft max activation function

The different activation functions are shown in Figure 1

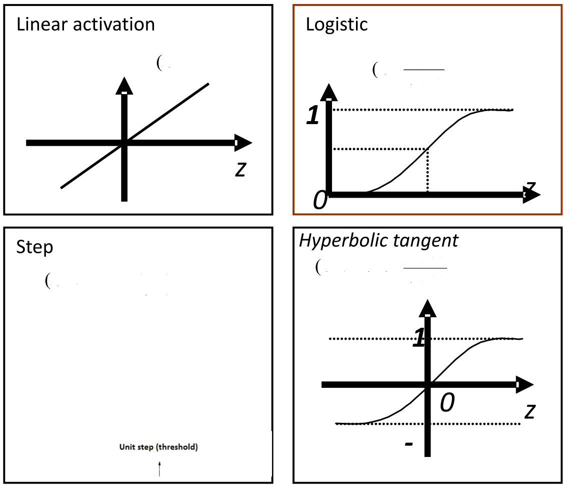

Figure 1 Different Activation functions

Neural networks with a linear activation function are not effective and are not suitable for multi layer networks. As the real world problems are non-linear, commonly non linear activations are used. The linear activation function normally is used in the hidden to output layer to transfer simply the computed values of hidden layer to the output layer. The Logistic activation function produces outputs in the range 0 to 1 when the neuron's net input varies from negative to positive.

The Tanh sigmoid activation function produces outputs in the range of 1 to -1 when the neuron's net input changes from positive to negative.

The sigmoid function has its output zero centred. So the Tanh sigmoid function is preferred in the place of the log sigmoid function.

1.2 Softmax activation function

Softmax function is also called an exponential but normalized function. It is a kind of logistic function that converts an X dimensional vector of real values into a same dimensioned vector of values, which fall in the range of 0 and 1 and also add up to 1 [1].

Let c be the number of output neurons (1)

$$y_i = \frac{e^{v_i}}{\sum_{i=1}^{c} e^{v_i}}, \quad i = 1, 2, \ldots, c$$

Where yi is the out put of ith neuron, v is the input.

$$\sum_{i=1}^{c} y_i = 1$$

It is clear that (2)

The softmax activation function is a smoothed version of a winnertake- all nonlinearity in which the maximum output is transformed to 1 (near 1), and all others reduced to 0 (near 0)

The softmax activation function, (Bridle 1990), assures that the outputs conform to the mathematical requirements of multi class and multivariate classification probabilities. Hence it is more suitable for multi class classification problems.

The softmax activation function is designed so that a return value is in the range (0,1) and the sum of all return values for a particular layer is 1.0. The softmax activation function is best explained by the following example.

For example, suppose three hidden-to-output sums are (2.0, -1.0, 4.0).

The scaling factor would be

Exp(2.0) + Exp(-1.0) + Exp(4.0) = 7.39 + 0.37 + 54.60 = 62.36.

Then the Softmax output values would be

Exp(2.0)/62.36, Exp(-1.0)/62.36, Exp(4.0)/62.36)= (0.12, 0.01, 0.87).

Notice that the final outputs are all between 0.0 and 1.0 and do in fact sum to 1.0, and further that the largest hidden-to-output sum (4.0) has the largest output/probability (0.87), and the relationship is similar for the second- and third-largest values

Hence the softmax function is employed for various multiclass probabilistic classification methods like, multi-class linear discriminant analysis, multi class naive Bayesian classifiers and multi-nomial logistic regression. Usually in an Elman Network the Sigmoid activation function is used in the input and hidden layers and linear activation function at the output layer. In this Research work a different activation function called the "Soft max" activation function is used which is more suitable for classification applications, especially multi class problems.

Furthermore Softmax function is differentiable and hence suitable for back propagation algorithm which is used in Elman networks.

2. Literature Review

Christopher M. Bishop [1] describes Softmax function and its usage in pattern recognition problems, using different methods including neural networks in his book Pattern Recognition and Machine Learning.

Paul Reverdy and N.E.Leonard [2] performed parametric estimation of decision making models which employ Softmax function.

Hao Peng et al. [3] presented a training method which incrementally trained the hierarchical softmax function for Neural Language models. They achieved a faster training speed of 30 times.

Binghui Chen et al. [4] proposed a Noisy softmax function for preventing the early saturation behaviour of the softmax function. They used it in Deep Learning Convolution Neural Networks. They conducted experiments using standard bench mark datasets and improved the performance of DCNNs.

Jun Mo Jeong et al. [5] proposed a novel method for correction of numerical errors due to exponential and logarithmic functions present in the soft-max function. They performed the transformation of the equations without affecting the features of existing softmax and cross entropy functions.

Dejian Yu [6] developed some novel aggregation operators under fuzzy environment with the help of Soft-max function. Multi-criteria decision making method using the Soft-max function based operators were developed and applied to flood disaster risk assessment problem.

Several Researchers performed Multi-class classi fication using various methods and for various applications.

The research conducted by Rajkumar Nallamuthu & Jacanathan Palanichamy [7] in the year 2015 examined different classification models constructed using thyroid dataset taken from machine learning repository, University of California. Nine effective feature subsets had been constructed. The subsets were tested with three most benchmarked algorithms namely C4.5, feed forward neural network and radial basis function network (RBFN) using various training-test partitions.

Dogantekin et al. [8] in 2010 developed an automatic diagnosis system based on thyroid gland ADSTG and obtained an accuracy of 97.67%

Dogantekin et al. [9] in 2011 developed an expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid diseases and obtained an accuracy of 91.86%

Feyzullah Temurtas [10] performed a comparative thyroid disease diagnosis using multilayer, probabilistic and learning vector quantization neural networks using the thyroid disease dataset taken from UCI machine learning database in 2012.

Keles and Keles[11] developed an Expert system for thyroid diseases Diagnosis (ESTDD) and obtained an accuracy of 95.33%.

Polat et al. [12] proposed a Novel hybrid method based on Artificial Immune recognition system (AIRS) with fuzzy weighted pre-processing for thyroid Disease diagnosis. He obtained an accuracy of 81% with AIRS technique and 85% with AIRS with fuzzy weighted pre-processing

Pasi (2004) performed classification applied to medical data sets especially the Thyroid data set using LDA, MLP, DIMLP, C4.5-1, C4.5-2 and C4.5 -3 techniques. The accuracy obtained ranged form 81.34 to 96.24.

Ishibuchi, H., Yamamoto, T. and Nakashima, T [14] carried out the classification process on different benchmark data sets extracted form UCI Machine Learning Data repository using the Hybridization of fuzzy Genetic Based Machine Learning (HGBML) approach.

Devaraj, D. and Ganeshkumar, P [15] used the Mixed Genetic Algorithm for the classification of data sets extracted form UCI Machine Learning Data repository

They built the fuzzy classifier using mixed genetic algorithm and performed the classification on the bench mark datasets successfully.

Dombi, J., & Gera, Z. [16] developed a squashing function based construction of a fuzzy rule based classifier in which the structure for the rules evolved by Genetic Algorithm and the fuzzy membership functions were fine tuned by gradient based optimization. The derivatives of membership functions were effectively computed by squashing function. They carried out the classification task using the benchmark datasets extracted form UCI Machine Learning Data repository. The results presented the efficiency of the classifier.

Serpen et al. [17] performed analysis of probabilistic potential function neural network classifiers using MLP, LVQ, RBFNN and PPFNN.

Nazri Mohammed Nawi et al. (2015) proposed a new meta-heuristic Cuckoo search (CS) algorithm for training of Elman Neural Network for classification applications.

They performed classification task on several benchmark datasets extracted form UCI Machine Learning Data repository. They established that the Hybrid Cuckoo search Elman Network Classifier performed better than the other algorithms.

3. Implementation and Results

The Elman neural network with Softmax activation function was simulated in MATLAB 2010.

The implementation methodology adopted in this work involved the following sub tasks

[1] Load the dataset (Normalize the data if required)

[2] Specify inputs and targets

[3] Split into training, validation and testing data. (70% of the data is used for training, 15% for validation and the remaining 15% is used for testing).

[4] Create the Neural network

[5] Configure network parameters (Specify Softmax activation function in the required layers)

[6] Train the network

[7] Simulate the network

[8] Get the output

[9] Measure the Metrics.

[10] Plot the required plots

[11] If the output obtained is not satisfactory then the architecture of the network is modified by modifying the hidden layers or the number of hidden layer neurons and the same method are adopted.

3.1 Data sets and Neural Network Architecture

The benchmark data sets were obtained from dataset obtained from the UCI Machine Learning Repository [19], which had a collection of databases, domains that were used by the machine learning community for the empirical analysis of machine learning algorithms.

The multi-class data sets used in this research work are :

1. Glass Identification Data set
2. Thyroid Disease Data set.

3.1.1 Glass Identification Data set

The Glass dataset was used for separating glass splinters in criminal investigation into 6(six) types of glasses namely float processed or non-float processed building windows, vehicle windows, containers, tableware, or head lamp (6 output classes). The data set was taken from UCI Repository or Machine Learning database [19] which consisted of 9 inputs and 6 outputs with 214 examples

Input attributes : 9

Output classes : 6

Number of instances : 214.

Hence the Elman Network had the following architecture.

Input Neurons - 9

Output Neurons - 6

Hidden layers -2

Hidden Layer Neurons 45,45

Activation function - Soft max in all layers

Glass Data set HGBML (GAGA) (MGA)Elman -
Softmax Accuracy 84.3 86.5 88.2 90.7

Training Algorithm - Scaled Conjugate Gradient Back propagation Algorithm.

The Classification Accuracy was found to be 90.7%. with Soft-max activation function and 89.67% with Sigmoid activation function which is shown in table 1.

Table 1. Classification Accuracy with different activations

Glass Identification Data	With Soft - max Activation	With Sigmoid Activation
Classification Accuracy	90.7%	89.67%

Figure 1 shows the comparison of accuracy with different activation functions

Fig 1. The comparison of accuracy with different activation functions

Further the results are compared with other methods available in literature. The table2 shows the performance (% accuracy) of glass dataset using different methods.

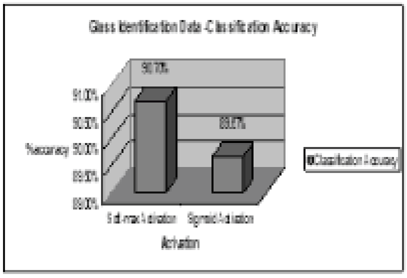

Table 2 Performance of GlassDataset

Glass Data set	HGBML	 (GAGA)	 (MGA)	Elman - Softmax
Accuracy	84.3	86.5	88.2	**90.7**

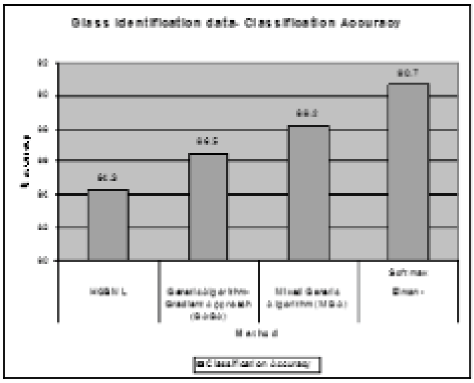

Fig 2 shows the performance plot of Glass dataset using different methods

Fig 2 . Performance plot of different methods

3.1.2 Thyroid Disease Data set

The thyroid data set has 21 input parameters, 3 output parameters, and 7200 data tuples. The three output classes are normal function, hyper function and hypo function of thyroid gland. Thyroid dataset is obtained from the UCI Machine Learning Repository [19].

There are 7200 patient records present in the data set.

The output classes are :

1. Normal, not hyperthyroid

2. Hyper function

3. Subnormal functioning

Hence the Elman Network has the following architecture.

Input Neurons - 21

Output Neurons - 3

Hidden layers -2

Hidden Layer Neurons 27, 27

Activation function - Soft max in all layers Training Algorithm - Scaled Conjugate Gradient Back propagation Algorithm.

The Classification Accuracy is found to be 97.8% with Soft-max activation function and 91.7% with Sigmoid activation function

as shown in table 3.

Table 3. Classification Accuracy with different activations

Thyroid Diagnosis Data	With Soft - max Activation	With Sigmoid Activation
Classification Accuracy	97.8%	91.7%

Figure 3 shows the comparison of accuracy with different activation functions

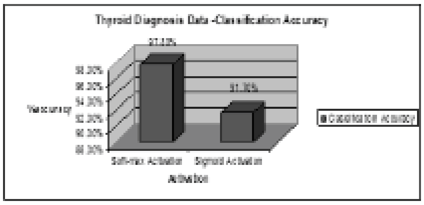

Fig 3. The comparison of accuracy with different activation functions

Further the results are compared with other methods available in literature. Tables 4 and 5 show the performance (% accuracy) of thyroid diagnosis dataset using different methods.

Table 4 Accuracy obtained by different algorithms for Thyroid data classification

Algorithms	Accuracy
ABC-BP [18]	93.28
ABC –BPLM [18]	91.66
ABC-NN [18]	88.18
BP-NN [18]	85.88
CSBP-ERN [18]	95.008
Proposed Elman — Softmax	**97.8**

Fig 4 and 5 show the performance plot of Thyroid Diagnosis classification using different methods

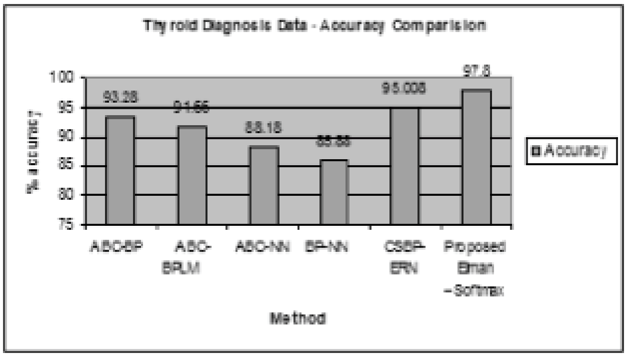

Fig 4 . Performance plot of different methods

Table 5 Accuracy obtained by different algorithms for Thyroid data classification

Method	Performance (% accuracy)
RBF [20]	79.08
CSFNN [20]	91.14
LVQ [17]	81.86
RBF [17]	72.09
C4.5-1[13]	93.26
C4.5-2 [13]	92.81
C4.5-3 [13]	92.94
MLP [13]	96.24
DIMLP [13]	94.86
ESTDDwith NEFCLASS-J [11]	95.33
AIRS [12]	81.00
AIRS with fuzzy [12]	85.00
ADSTG [8]	97.67
GDA-WSVM [9]	91.86
Improved F - score with MLP [7]	93.49
Proposed Elman Network	97.8

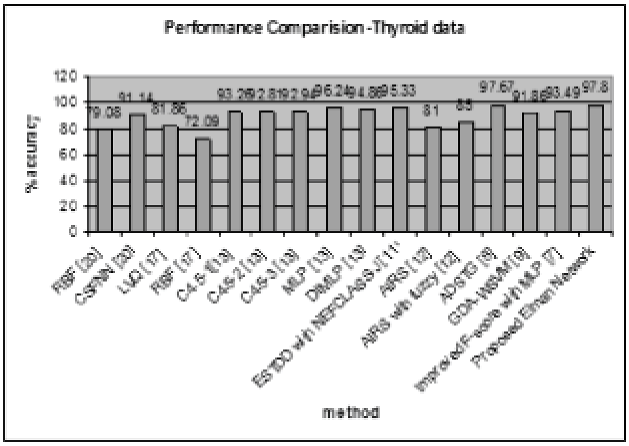

Fig 5.. Performance plot of different methods

4. Discussions

The Elman neural network with Soft max activation function is implemented by training and testing the bench mark datasets from UCI machine learning Repository. The classification accuracy is improved compared to the accuracy obtained with Sigmoid activation function. Further compared to the accuracies of other methods in the literature, the proposed Elman network with Softmax activation offers better accuracy.

The simulations are carried out using different training algorithms and the Scaled Conjugate Gradient Back propagation Algorithm shows better results.

Further, there are no definite rules for fixing the number of Hidden layers and the number of neurons in each hidden layer. Several trials are run with different number of hidden layers and hidden layer neurons and the setup which offers better results were considered.

5. Conclusion

In this research work the effectiveness of Soft-max function as an activation function for Neural Networks is analyzed. An Elman network with Softmax activation function is simulated which is trained and tested using two bench mark data sets obtained from UCI machine learning repository. One of the data sets, the Glass identification sets has a large number of output classes (6 output classes) and nine input attributes. The other dataset namely the Thyroid Diagnosis data set has three output classes and 21 input attributes and is a large data set with 7200 data rows. The proposed Softmax activated Elman Neural Network performed a better classification and showed a better performance than the Sigmoid-activated Elman neural network and also showed a better performance than the existing methods.

References :

1. Christopher M.Bishop, "Pattern Recognition and machine learning", Springer, 2006, ISBN - 10: 0387-31073-8.
2. Paul Reverdy and Naomi Ehrich Leonard," Parameter Estimation in Softmax Decision Making Models with Linear Objective Functions" , IEEE Transactions in Automation Science & Engineering, Vol 13, Jan 2016, pages 54-67.
3. Hao Peng, Jianxin Li, Yangqiu Song and Yaopeng Lin, " Incrementally Learning the Hierarchical Softmax function for Neural Language Models", Proceedings of Thirty first AAAI conference on Artificial Itelligence 2017, pages 3267-3273.
4. Binghui Chen, Weihong Deng, Junping Du, "Noisy Softmax : Improving the Generalization Ability of DCNN via Postponing the Early Softmax Saturation,open access CVPR paper , Computer Vision Foundation, pages 5371-5382.
5. Jun Mo Jeong, Se Jin Choi and Chinyong Kim , " Correcting Method for of Numerical errors of Soft-max and Cross-entropy according to CNN's output value",Asia-Pacific Journal of Neural Networks and applications, vol 1 No. 1(2017) pp 15-20.
6. Dejian Yu, "Softmax function based intuitionistic fuzzy multi-criteria decision making and applications", Operational Research an international journal ,July 2016,Volume 16, issue 2, pp 327-348.
7. Rajkumar Nallamuthu & Jacanathan

Palanichamy(2015), "Optimized construction of various classification models for the diagnosis of thyroid problems in human beings" , Kuwait Journal of Science, 42 (2) pp. 189-205, 2015.

8. Dogantekin, E., Dogantekin, A. & Avci, D. 'An automatic diagnosis system based on thyroid gland: ADSTG', Expert Systems with Applications, 2010, 37(9) pp.6368-6372.

9. Dogantekin, E., Dogantekin, A. & Avci, D. An expert system based on Generalized Discriminant Analysis and Wavelet Support Vector Machine for diagnosis of thyroid diseases, Expert Systems with Applications, 38:146-150,2011

10. Feyzullah Temurtas, A comparative study on thyroid disease diagnosis using neural networks, Expert Systems with Applications: An International Journal, 2012, 36: 944-949.

11. Keles, A. & Keles, A. ESTDD: Expert system for thyroid diseases Diagnosis, Expert Systems with Applications. 34: 242-246,2008.

12. Polat, K., Sahan, S. & Gunes, S., "A Novel hybrid method based on Artificial Immune recognition system (AIRS) with fuzzy weighted preprocessing for thyroid Disease diagnosis", Expert Systems with Applications, 32: 1141- 1147,2007.

13. Pasi, L, "Similarity classifier applied to medical data sets," 10 sivua, Fuzziness in Finland'04 in: International conference on soft computing, Helsinki, Finland &Gulf of Finland & Tallinn, Estonia ,2004

14. Ishibuchi, H., Yamamoto, T. and Nakashima, T. Hybridization of fuzzy GBML approaches for pattern classification problems, IEEE Transaction on System Man and Cybernetics Part B, Vol.35, No.2, pp.359-365, 2005.

15. Devaraj, D. and Ganeshkumar, P. Mixed Genetic Algorithm approach for Fuzzy Classifier Design, International Journal of Computational Intelligence and Applications, Vol.9, No.1, pp.49-67, 2010.

16. Dombi, J., & Gera, Z. Rule based fuzzy classification using squashing functions. Journal of Intelligent and Fuzzy Systems, 19(1), 3-8,2008.

17. Serpen, G., Jiang, H. & Allred, L. Performance analysis of probabilistic potential function neural network classifier in: Proceedings of artificial neural networks in engineering conference, St.Louis, MO. 7: 471 - 76,1997.

18. Nazri Mohammed Nawi, Abdullah Khan, Rehman. M.Z, Haruna Chiroma and Tutut Herawan, "Weight Optimization in Recurrent Neural Networks with Hybrid Metaheuristic Cuckoo Search Teechniques for Data Classification" (2015), Mathematical Problems in Engineering , Hindawi Publishing Corporation, volume 2015, open access article.

19. UCI Machine Learning Repository [http://archive. ics. uci. edu/ml]. Irvine, CA: University of California. School of Information and Computer Science.

20. Ozyılmaz, L. & Yıldırım, T. Diagnosis of thyroid disease using artificial neural Network methods in: Proceedings of ICONIP'02 ninth international conference on neural information processing, Orchid Country Club, Singapore, 2002.