ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

Classification System to Predict Protein Localization Sites

Dr. Kalpana Chansoria [1]

Sr. Lecturer, Government Polytechnic College, Mandla, Madhya Pradesh, India[1]

ABSTRACT: Many Classification Techniques have been used to predict Protein Localization sites in E Coli. Neural Networks are soft computing tool which can be used for various computational tasks such as classification, predictions, function fitting etc. Proposed work tries to predict Protein Localization Sites using ANN classifier.

KEYWORDS: E Coli, Artificial Neural Networks, classification, protein localization site.

I. INTRODUCTION

E Coli. is a parasitic species which creates protein localization by many biological changes. E Coli database is a benchmark database for prediction. Artificial Neural Networks are promptly used for classification purposes. It uses weighted sum of inputs passing through a nonlinear activation function enabling it to classify and predict complex patterns in input data.

II. RELATED WORK

In[1] Zhi-Vassilis Athitsos and Stan Sclaroff. Have used unsupervised neural network for classification, In[2] Charles X. Ling and Qiang Yang and Jianning Wang and Shichao Zhang have done classification with the help of Bayes classifier, Aik Choon an and David Gilbert [4] have shown results of different supervised learning networks. Huajie Zhang and Charles X. Ling[6] have proposed a new algorithm for classification through bayes classifier. Vyas et.al. [7] have done multi class classification of eukaryotic protein localization site using FF-ANN.

III. PROTEIN LOCALIZATION PREDICTION

To predict protein localization sites we propose to use feed forward neural network. EColi bacteria induces protein localization because of various secretions and other biological activities. Database for the same has been taken from the UCI machine learning repository. We have used offline trainlm algorithm to train our ANN. Neural network network tool box (nftool) (MATLAB) is used for the same.

3.1 E Coli Database

This database was created by Mr Kenta Nakai is a prime source to experiment classification of protein localization phenomena. This database has originally 336 instances of 8 attributes with no missing values. First attribute denotes only a sequence number of SWISS-PROT database and has no relevance in classification. Other attributes which are used in our experiment are as follows

1. mcg: McGeoch's method for signal sequence recognition.
2. gvh: von Heijne's method for signal sequence recognition.
3. lip: von Heijne's Signal Peptidase II consensus sequence score. Binary attribute.
4. chg: Presence of charge on N-terminus of predicted lipoproteins.Binary attribute.
5. aac: score of discriminant analysis of the amino acid content ofouter membrane and periplasmic proteins.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

6. alm1: score of the ALOM membrane spanning region prediction program.
7. alm2: score of ALOM program after excluding putative cleavable signal regions from the sequence.

Where as The class is the localization site as depicted below.-

i. cp (cytoplasm)
ii. im (inner membrane without signal sequence)
iii. pp (perisplasm)
iv. imU (inner membrane, uncleavable signal sequence)
v. om (outer membrane)
vi. omL (outer membrane lipoprotein)
vii. imL (inner membrane lipoprotein)
viii. imS (inner membrane, cleavable signal sequence)

3.2 CLASSIFICATION

We have used nftool (A MATLAB tool box) for classification of above data set. There are a total of 8 classes present in sample space. We have used 'ONE HOT' notation to represent each classes. Avoiding over fitting by network while preserving generalizability we have found 3 neurons for hidden layer through repetition of experiment on various configuration. Structure for the same is shown in fig.1.

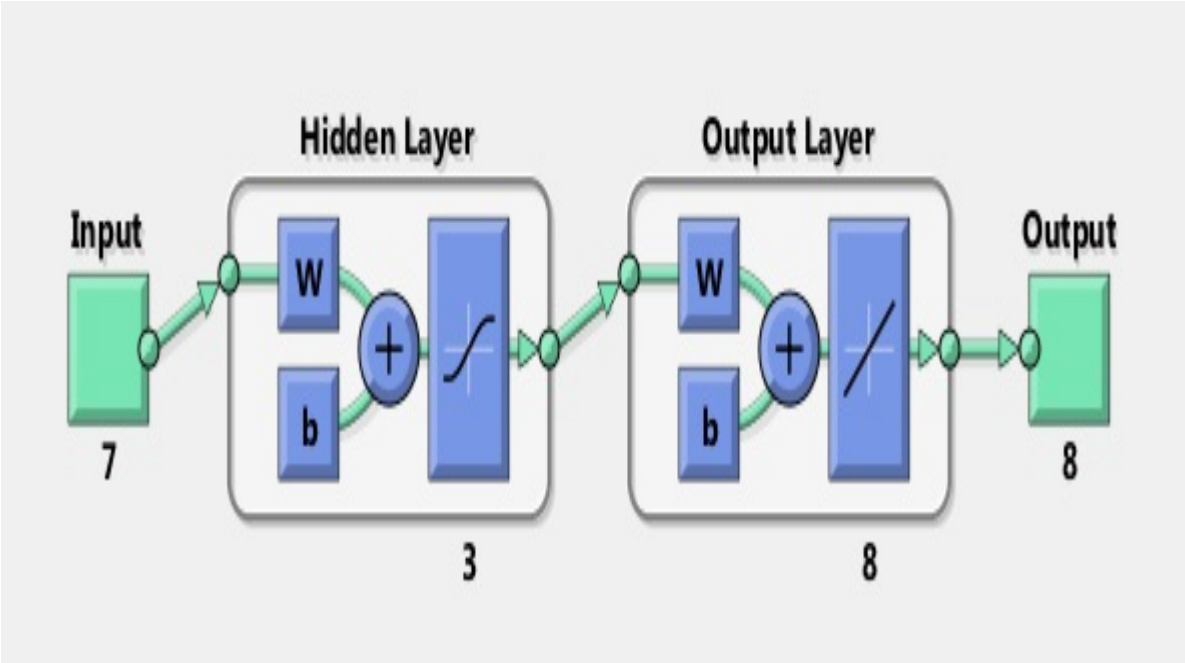

fig1: Proposed Structure of Artificial Neural Network

IV. EXPERIMENTAL RESULTS

MATLAB version R2013a is used in our experiment. The E Coli Database (downloaded from the UCI repository, www.ics.uci.edu) has 336 X 8 matrix. After removing un necessary attribute a 336 X 7 matrix is prepared, as input data and 336 X 8 matrix is prepared as output data. Out of these 336 samples, 70% sample were used for training purpose, 15% for validation and 15% for testing.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

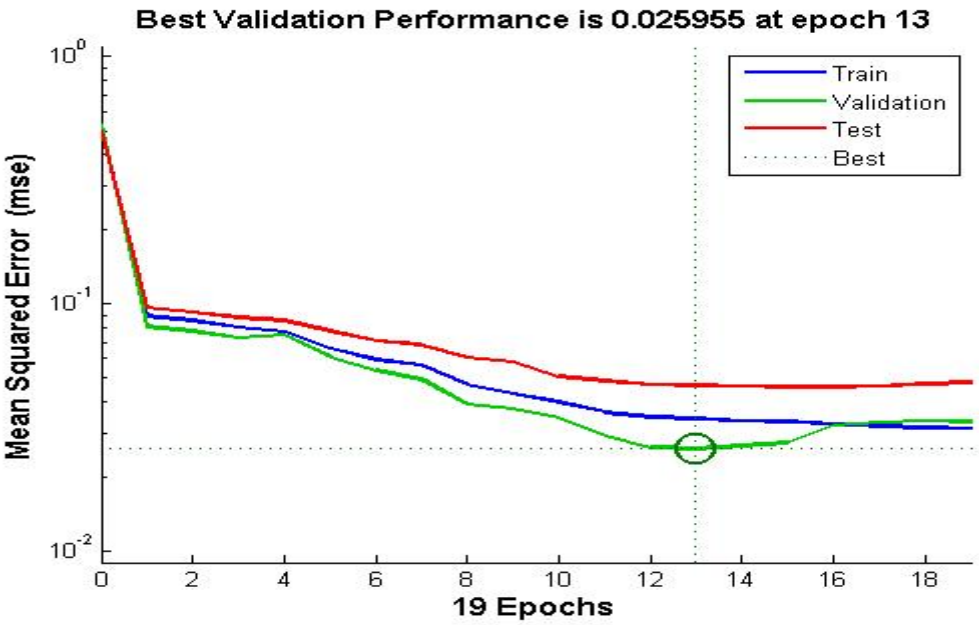

Fig 2.1: Validation Performance

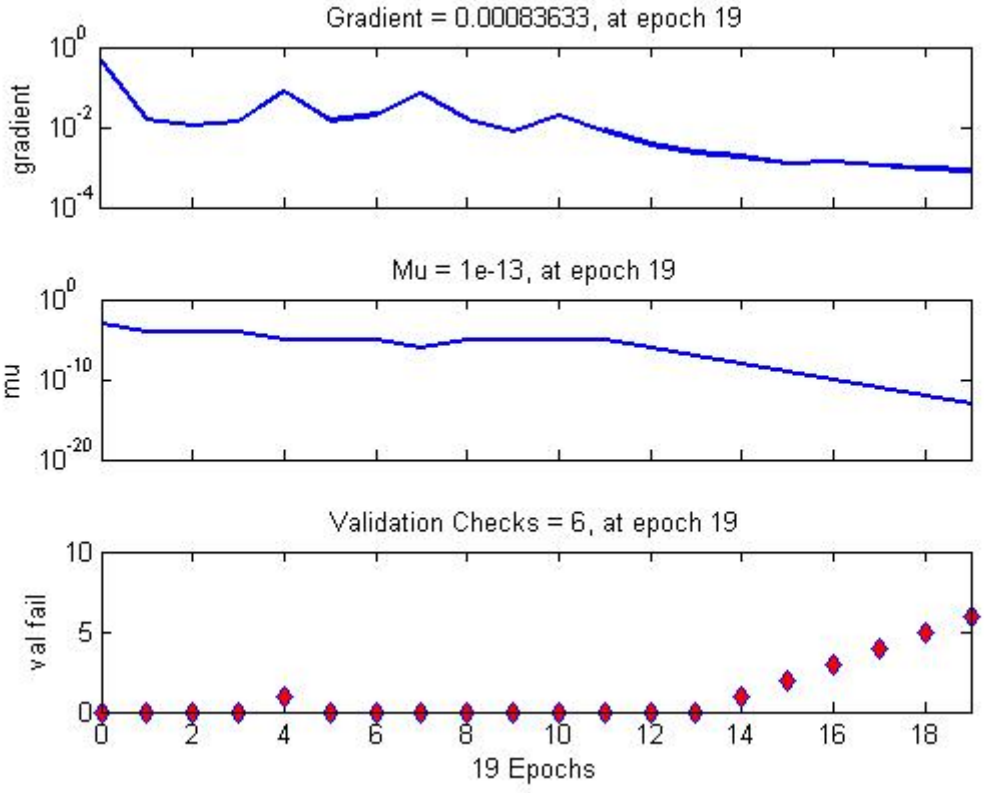

Fig 2.2 Gradient Analysis

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

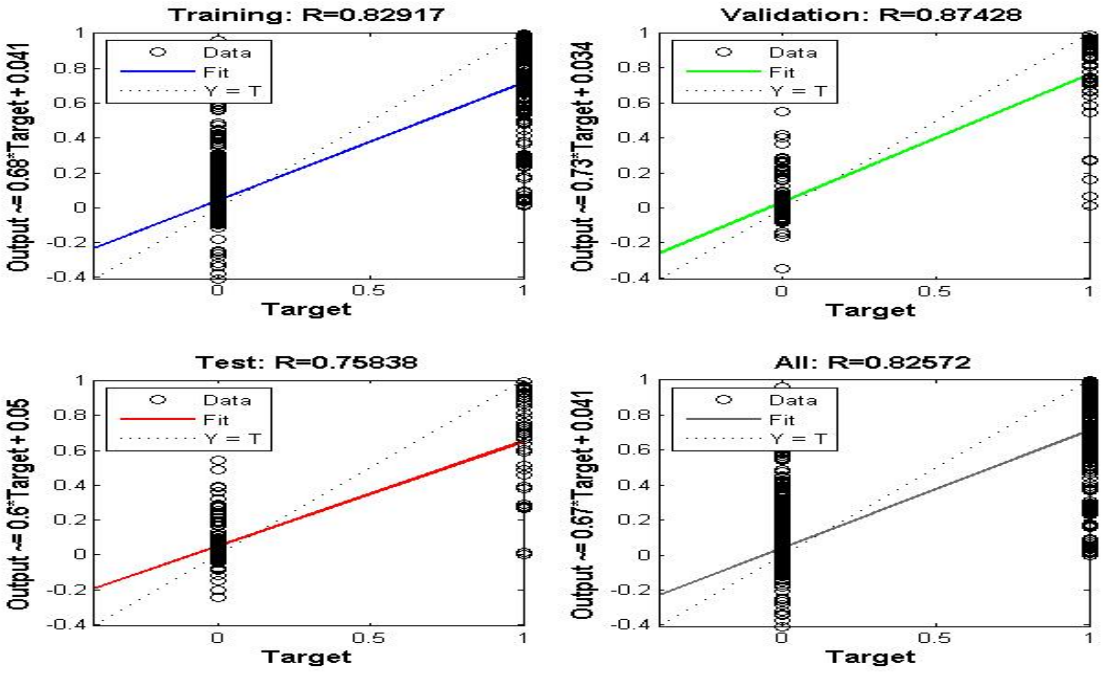

Fig 2.3: Regression Analysis

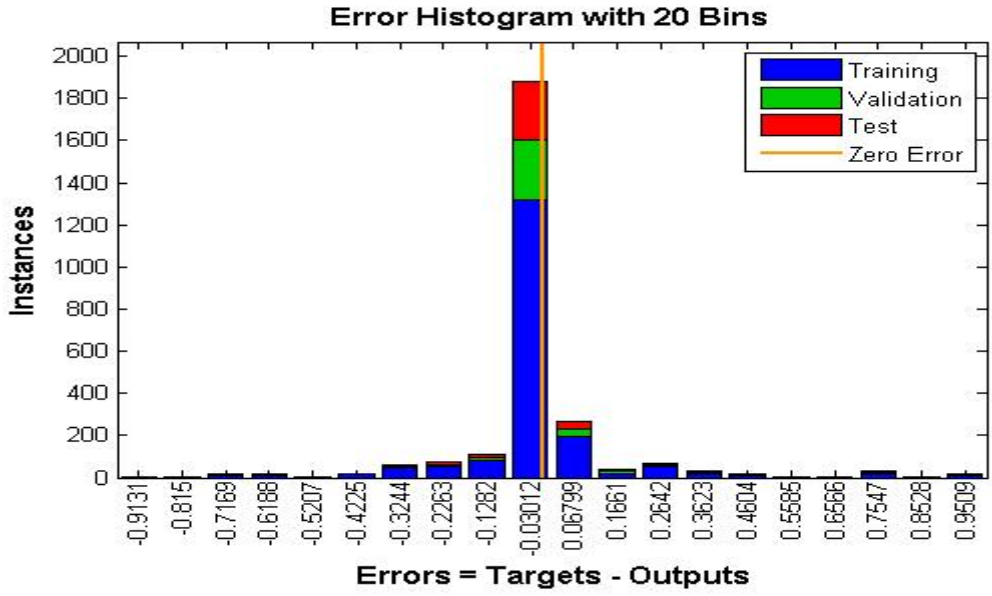

Fig 2.4: Error Histogram

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Visit: www.ijirset.com

Vol. 7, Issue 10, October 2018

V. CONCLUSION

MLFANN successfully predicted the protein localization sites. As shown in result 2.1 best genralizable solution is found at 13th Epoch while training error keeps decreasing. Testing result also support this fact. So, from above results and discussion it can be concluded that an artificial neural network can be effectively used to predict and in turn classify our data set

REFERENCES

[1] Zhi-Vassilis Athitsos and Stan Sclaroff. Boosting Nearest Neighbor Classifiers for Multiclass Recognition. Boston University Computer Science Tech. Report No, 2004-006. 2004.
[2] Charles X. Ling and Qiang Yang and Jianning Wang and Shichao Zhang. Decision trees with minimal costs. ICML. 2004.
[3] Xiaoyong Chai and Li Deng and Qiang Yang and Charles X. Ling. Test-Cost Sensitive Naive Bayes Classification. ICDM. 2004.
[4] Aik Choon an and David Gilbert. An Empirical Comparison of Supervised Machine Learning Techniques in Bioinformatics. APBC. 2003.
[5] Mukund Deshpande and George Karypis. Evaluation of Techniques for Classifying Biological Sequences. PAKDD. 2002.
[6] Huajie Zhang and Charles X. Ling. An Improved Learning Algorithm for Augmented Naive Bayes. PAKDD. 2001.
[7] Shrikant Vyas et.al. Predicting Protein Localization Sites in Eukaryotic Cells Using Yeast Data Base through Feed forward Neural Network. IRJES. 3(12): December, 2014