ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

Predicting Protein Localization Sites in Eukaryotic Cells Using Yeast Data Base through Feedforward Neural Network

Shrikant Vyas, Dipti Upadhyay*
Department of Cyber Law And Information Technology, Barkatullah University, Bhopal, Madhya Pradesh, India
*Department of Biomedical Engineering, Barkatullah University, Bhopal, Madhya Pradesh, India

Abstract
The prediction on the basis of individual features plays an important role in behavioral sciences. Current statistical methods do not always yield satisfactory answers. A Feed Forward Artificial Neural Network is a statistical tool which can be used as a computer model inspired by the performance of the Human Brain. It views as in the set of artificial neurons that are interconnected with the other neurons. The primary aim of this paper is to demonstrate the process of prediction of protein localization sites in eukaryotic cells using a yeast database with the leap of a feed forward neural network. This paper is related to the use of feed forward neural networks towards the prediction in the condition where a dataset is linearly inseparable with the help of error back propagation algorithm (ebpa

Keywords: yeat dataset, Feed Forword Artificial Neural Networks, linear inseperability, ebpa

Introduction
Prediction is the organization of data into predefined groups. It comes under supervised learning method as the classes are determined before examining the data. In behavioral sciences, as well as in most biological sciences, statistical analyses using traditional algorithms do not always lead to a satisfactory solution, particularly in classification analysis. Current classification methods rely on parametric or non-parametric multivariate analyses: Discriminate analysis, cluster analyses, etc. Many works on multi-label learning deal with text categorization problems. Schapire and Singer [3] Proposed the famous BoosTexter approach by extending from the popular ensemble learning Method AdaBoost [7], where a set of weights over all instance-label pairs are maintained and Those pairs that are hard (easy) to predict correctly will get incrementally higher (lower) weights. McCallum [9] and Ueda and Saito [8] respectively proposed two multi-label text categorization Approaches by assuming generative models based on text frequencies. McCallum [9] assumed that A mixture probabilistic model (one mixture component per category) is assumed to generate each Document and utilized EM [6] algorithm to learn the mixture weights and the word distributions In each mixture component. Ueda and Saito [8] presented two types of probabilistic generative Models for multi-label text called parametric mixture models (Pmm1, Pmm2). They assumed that multi-label text has a mixture of characteristic words appearing in single-label text that belongs to Each category of the multi-categories. Gao et al. [5] Proposed a maximal figure-of-merit (MFoM) Approach [1] for multi-label text categorization, where classifier parameters are incorporated into A continuous and differentiable function which simulates specific performance metrics, and then Learned by optimizing the specified function.All approaches to performing classification assume some knowledge of the data. Usually, a training set is used to develop the specific parameters required. Pattern classification aims to build a function that maps the input feature space to an output space of two or more than two classes.

Materials and methods
We have used a feed forward neural network in order to classify the yeast data set. The yeast data set is one of the benchmark data sets used to demonstrate the approach for prediction problems. We have taken this dataset from the UCI machine learning

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

repository. We have used EBPA algorithm to train our ANN. Since the gradient decent methodology requires a differentiable activation function, we have used log sigmoid function. Log sigmoid function is an S – shaped function having a range between -1 to 1.Because of it, we have modified the classes of target data set ranging from -1 to +1 since it's a nine class classification problem. Matlab neural network tool box (nntool) is used to do the necessary classification task.

Yeast dataset

The yeast dataset is a 9 class classification problem which we have obtained from use machine learning repository. Development of the data set is explained in [2] and [4]. The Yeast Database Contains the following properties:

1. Sequence Name: Accession number for the SWISS -PROT database.
2. mcg: McGeoch's method for signal sequence recognition.
3. gvh: von Heijne's method for signal sequence recognition.
4. alm: Score of the ALOM membrane spanning region prediction program.
5. mit: Score of discriminant analysis of the amino acid content of the N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins.
6. erl: Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute.
7. Pox: Peroxisomal targeting signal in the C-terminus.
8. vac: Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins.
9. nuc: Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins.

The first attribute which is only an accession number has been omitted since there is no significance of that attribute in prediction on the other hand, we have converted target values to numeric values ranging from 0 to +1 in such a way that the classes have a maximum distance in terms of numeric values.

Classification

An ANN is an information-processing system that is based on the simulation the human cognition process. ANNs consist of many computational neural units connected to each other.

The advantages of Neural Networks for classification are:

- Neural Networks are more robust because of the weights
- The Neural Networks improve its performance by learning. This may continue even after the training set has been applied.
- The use of Neural Networks can be parallelized as specified above for better performance.
- There is a low error rate and thus a high degree of accuracy once the appropriate training has been performed.
- Neural Networks are more robust in noisy environment.

In ANN, knowledge about the problem is distributed in neurons and connection weights of links between neurons. The neural network has to be prepared to adjust the connection weights and biases in society to create the desired mapping. ANNs are particularly useful for complex pattern recognition and classification tasks. The capability of learning from examples, the ability to reproduce arbitrary non-linear functions of input, and the highly parallel and regular structure of ANNs make them especially suitable for pattern classification problems. The most commonly used training algorithms is the back propagation (BP) algorithm with gradient descent, which is used in this work also. This algorithm is based on the adjustment of the weights of the connections of the network to minimize error. The error is calculated by comparing obtaining outputs with expected outputs of known inputs. This error is then backward propagated until the first layer and the weights are then adjusted. This process occurs over and over as the weights are continually adjusted. The set of data which enables the training is called "training set". During the training of a network, the same set of data is processed many times until reaching an acceptable error, or reaching the maximum number of iterations.

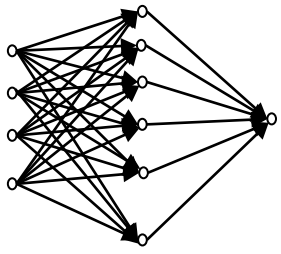

Figure 1: Proposed ANN Architecture

Results and discussion

Results

The MATLAB version used is R2013a. The YEAST dataset (downloaded from the UCI repository,

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

www.ics.uci.edu, which is a 1484×9 matrix, is first processed to remove accession number and then the remaining 1484*8 matrix taken as the input data. Out of these 1484 samples, 70% sample was used for training, 15% for validation and 15% for testing. Under supervised learning, the target of different inputs are in a random order to ensure proper training. The network architecture taken was 8×200×1, i.e, the input layer has 8 nodes, the hidden layer has 200 nodes and the output layer has 1 node.

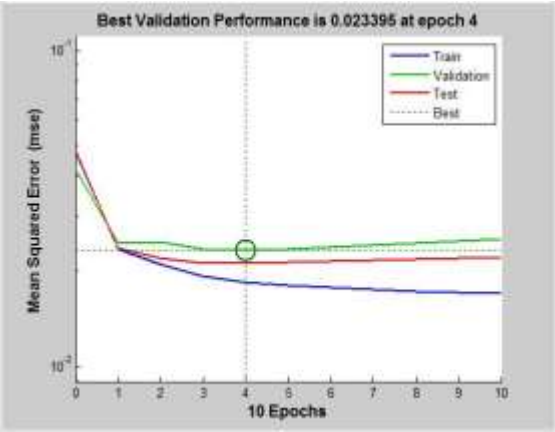

Figure 2.1

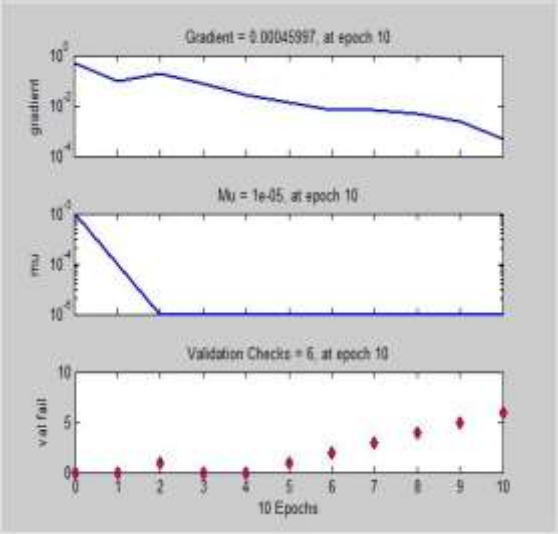

Figure 2.2

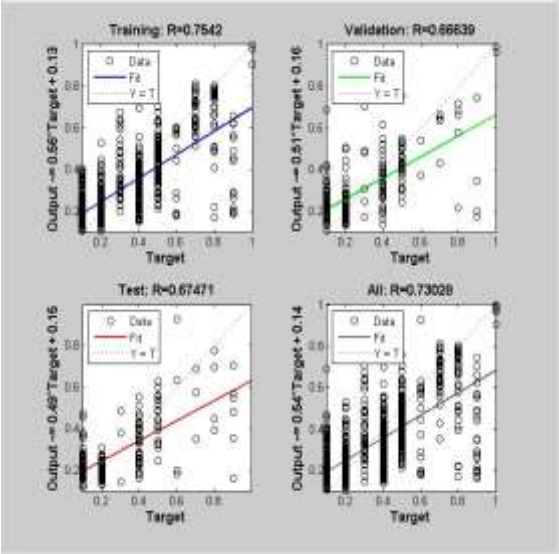

Figure 2.3

Conclusion

The Multi Layer Feed Forward Neural network gives us a satisfactory result, because it is able to predict the nine different types of protenes of 1484 instances with just few errors for the other one.From the graphs we observe that Back propagation Algorithm gives the best accuracy. Best performance was obtained on 4th epoch as on validity cheack graph it shows less generelization after 4th epoch. From the above results, graphs and discussion, it is concluded that Multi Layer Feed Forward Neural Network (MLFF) is faster in terms of learning speed and gave a good accuracy, i.e., has the best trade-off between speed and accuracy. So, for faster and accurate classification, Multi Layer Feed Forward Neural Networks can be used in many pattern classification problems.

Acknowledgements

The authors are thankful to the management of Barkatullah University for providing the necessary facilities to undertake the above work.

References

[1] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua. A maximal figure-of-merit learning approach to text categorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 174–181, Toronto, Canada, 2003.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 2.114

[2] Kenta Nakai & Minoru Kanehisa "A Knowledge Base for Predicting Protein Localization Sites in Eukaryotic Cells", Genomics 14:897-911, 1992.
[3] R. E. Schapire and Y. Singer. Boostexter: a boosting-based system for text categorization. *Machine Learning*, 39(2/3):135–168, 2000.
[4] Kenta Nakai & Minoru Kanehisa "Expert Sytem for Predicting Protein Localization Sites in Gram-Negative Bacteria", PROTEINS: Structure, Function, and Genetics 11:95-110, 1991.
[5] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua. A MFoM learning approach to robust multiclass multilabel text categorization. In *Proceedings of the 21st International Conference on Machine Learning*, pages 329–336, Banff, Canada, 2004.
[6] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistics Society -B*, 39(1):1–38, 1977.
[7] Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.
[8] N. Ueda and K. Saito. Parametric mixture models for multi-label text. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 721–728. MIT Press, Cambridge, MA, 2003.
[9] A. McCallum. Multi-label text classification with a mixture model trained by EM. In *Working Notes of the AAAI'99 Workshop on Text Learning*, Orlando, FL, 1999.