Republic of Iraq
Ministry of Higher Education and Scientific Research
Baghdad University - College of Science
Computer Science Department

Iris Flowers Classification Using Neural Network

A Project **BY**
Assist. Lec. Sura Abed Sarab

June 2020 **Shawal 1441**

Table of Contents

List of Tables

List of Figures

List of Abbreviations

LR	Learning Rate
ANN	Artificial Neural Network
PCA	Principle component analysis
ML	Machine Learning
EV	Eigenvalue
LR	Learning rate
NH	Number of hidden neurons
BPNN	Back Propagation Neural Network
FFNN	Feed Forward Neural Network
AI	Artificial Intelligence
MLFF	Multi-Layer Feed Forward

Abstract

Classification is a machine learning technique used to predict group membership for data instances. To simplify the problem of classification neural networks are being introduced. This report focuses on IRIS plant classification using Neural Network. The problem concerns the identification of IRIS plant species on the basis of plant attribute measurements. Classification of IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS plant and how the prediction was made from analyzing the pattern to form the class of IRIS plant. By using this pattern and classification, in future upcoming years the unknown data can be predicted more precisely. Artificial neural networks have been successfully applied to problems in pattern classification, function approximations, optimization, and associative memories. In this work, Multilayer feed-forward networks are trained using back propagation learning algorithm. The experiential results show the minimum error rate was *0.01067* with training time of *0.691* millisecond, and the number of hidden neurons was 4.

CHAPTER ONE

Introduction

1.1 Overview

Bioinformatics is a promising and novel research area in the 21st century. This field is data-driven and aims at the understanding of relationships and gaining knowledge in biology. In order to extract this knowledge encoded in biological data, advanced computational technologies, algorithms, and tools need to be used. Basic problems in bioinformatics like protein structure prediction, multiple alignments of sequences, phylogenic inferences, etc. Are inherently non-deterministic polynomial-time hard in nature. To solve these kinds of problems artificial intelligence (AI) methods offer a powerful and efficient approach. Researchers have used AI techniques like Artificial Neural Networks (ANN), Fuzzy Logic, Genetic Algorithms, and Support Vector Machines to solve problems in bioinformatics. Artificial Neural Networks is one of the AI techniques commonly in use because of its ability to capture and represent complex input and output relationships among data. The purpose of this paper is to provide an overall understanding of ANN and its place in bioinformatics to a newcomer to the field.

Classification is one of the major data mining processes which maps data into predefined groups. It comes under supervised learning method as the classes are determined before examining the data. 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. Neural Networks (NN) are an effective tool in the field of pattern classification. Neural networks are simplified models of the biological nervous systems. An NN can be said to be a data processing system, consisting of a large number of simple, highly interconnected processing elements (artificial neurons), in an architecture inspired by the structure of the cerebral cortex of the brain. The interconnected neural computing elements have the quality to learn and thereby acquire knowledge and make it available for use. NN is an effective tool in the field of pattern classification. This project is related to the use of multi-layer feed-forward neural networks (MLFF) and backpropagation

algorithm towards the identification of IRIS flowers based on the following measurements: sepal length, sepal width, petal length, and petal width. A variety of constructive neural-network learning algorithms have been proposed for solving the general function approximation problem. The traditional BP algorithm typically follows a greedy strategy wherein each new neuron added to the network is trained to minimize the residual error as much as possible. This report also contains an analysis of the performance results of backpropagation neural networks with various numbers of hidden layer neurons and the different number of epochs.

1.2 Related Work

There are some experts that understand the IRIS dataset very well. There are few experts that have done research on this dataset. The researchers have mentioned that there isn't any missing value found in any attribute of this data set.

1- (Satchidananda Dehuri and Sung-Bae Cho) presented a new hybrid learning scheme for Chebyshev functional link neural network (CFLNN); and suggest possible remedies and guidelines for practical applications in data mining. The proposed learning scheme for CFLNN in classification is validated by an extensive simulation study. Comprehensive performance comparisons with a number of existing methods are also presented.

2- Saito and Nakano proposed a medical diagnosis expert system based on a multilayer ANN in. They treated the network as a black box and used it only to observe the effects on the network output caused by change the inputs.

3- Fernández-Redondo M. and Hernández-Espinosa C. reviewed two very different types of input selection methods: the first one is based on the analysis of a trained multilayer feed forward neural network (MFNN) and the second ones is based on an analysis of the training set. They also present a methodology that allows experimentally evaluating and comparing feature selection methods.

4- Two methods for extracting rules from ANN are described by Towell and Shavlik in the first method is the subset algorithm which searches for subsets of connections to a

node whose summed weight exceeds the bias of that node. The major problem with subset algorithms is that the cost of finding all subsets increases as the size of the ANNs increases. The second method, the MofN algorithm is an improvement of the subset method that is designed to explicitly search for M-of-N rules from knowledge based ANNs. Instead of considering an ANN connection, groups of connections are checked for their contribution to the activation of a node, which is done by clustering the ANN connections.

1.3 Project objectives

The current study aims to identify the type of iris flowers by using the dataset that prepared in advance way by the expert biologists to study the flower types through some measurements and statistics for each type using data mining techniques and neural network classifiers.

1.4 Project organization

This report is organized as the following manner: Chapter one presents a general overview and related work of our project. Chapter two presents the materials and methodology and how the researcher classifies flower types. Chapter 3 is based on the details of programming language techniques and source code. Chapter 4 gives the experimental results that obtained from the research. Finally, in chapter 5 conclusion and possible directions for future work are given.

CHAPTER TWO

Problem Statement and Methodology

Problem statement and methodology

This chapter will explain the proposed method in detail and how the project is concerned to classify each type in the right way.

2.1 Problem Statement

The real problem in this study is how to achieve a new way to classify Iris flowers and identify their type to study their behavior and help biologists with new ML techniques.

2.2 Pattern Recognition

Pattern Recognition is a fundamental human intelligence. In our daily life, we always do 'pattern recognition', for instance, we recognize images. Basically, pattern recognition refers to analyzing information and identifying for any kind of forms of visual or phenomenon information. Pattern recognition can describe, recognize, classify and explain the objects or the visual information.

As machine learning, pattern recognition, can be treated as two different classification methods: supervised classification and unsupervised classification. They are quite similar to supervised learning and unsupervised learning. As supervised classification needs a teacher that gives the category of samples, the unsupervised classification is doing it the other way around. Pattern recognition is related to statistics, psychology, linguistics, computer science, biology and so on. It plays an important role in Artificial Intelligence and image processing.

2.3 The Material

One of the most popular and best known databases of the neural network application is the IRIS plant data set which is obtained from UCI Machine Learning Repository and created by R.A. Fisher while donated by Michael Marshall (MARSHALL%PLU@io.arc.nasa.gov) on July, 1988 The IRIS dataset classifies three different classes of IRIS plant by performing

pattern classification [8]. The IRIS data set includes three classes of 50 objects each, where each class refers to a type of IRIS plant. The attributed that already been predicted belongs to the class of IRIS plant. The list of attributes present in the IRIS can be described as categorical, nominal and continuous.

The experts have mentioned that there isn't any missing value found in any attribute of this data set. The data set is complete. This project makes use of the well-known IRIS dataset, which refers to three classes of 50 instances each, where each class refers to a type of IRIS plant. The first of the classes is linearly distinguishable from the remaining two, with the second two not being linearly separable from each other. The 150 instances, which are equally separated between the three classes, contain the following four numeric attributes:

1- **sepal length**
2- **sepal width**
3- **petal length**
4- **petal width**

the fifth attribute is the predictive attributes which is the class attribute that means each instance also includes an identifying class name, each of which is one of the following: IRIS **Setosa**, IRIS **Versicolour**, or IRIS **Virginica**.

The expectation from mining IRIS data set would be discovering patterns from examining petal and sepal size of the IRIS plant and how the prediction was made from analyzing the pattern to form the class of IRIS plant. By using this pattern and classification, the unknown data can be predicted more precisely in upcoming years. It is very clearly stated that the type of relationship that being mined using IRIS dataset would be a classification model. This can classify the type of IRIS plant by examining the sizes of petal and sepal. Sepal width has positive relationship with Sepal length and petal width has positive relationship with petal length. This pattern is identified with bare eyes or without using any tools and formulas. It is realized that the petal width is always smaller than petal length and sepal width also smaller than sepal length.

2.4 Methodology

The proposed method of this work by the researcher includes the following steps:

2.4.1 Preprocessing

Initially the dataset pre-processing was done by isolating the attribute part that have the features from the label part and cleaning the data by removing null or missing values.

The dataset is shuffled to ensure that all cases will be trained on the neural network and tested successfully.

In addition, the dataset has been configured to be an input to the neural network, by normalize it and make its values between zero and one to reduce the model over-fitting by the following equation.

$$Z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

2.4.2 Artificial Neural Network

Neural networks are composed of simple elements operating in parallel. The neuron model shown in Figure 2.1. is the one that widely used in artificial neural networks with some minor modifications on it.

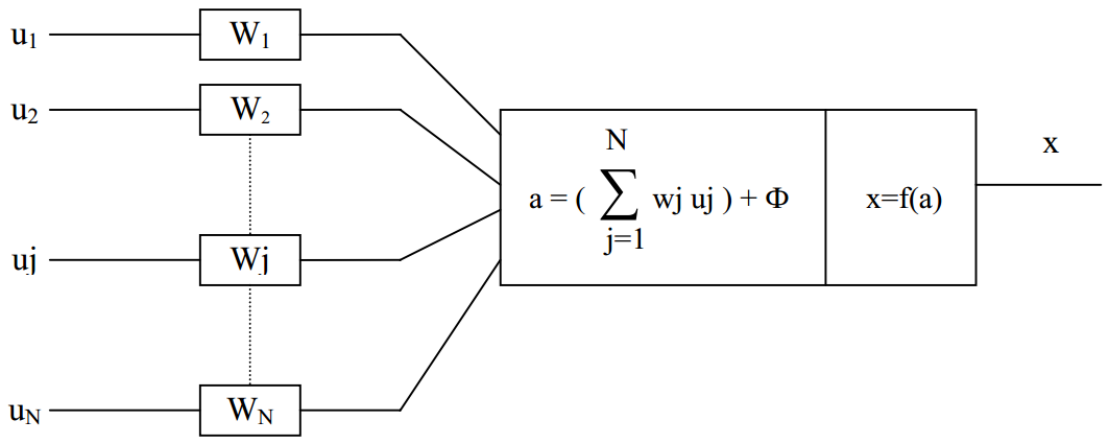


Figure 2.2 Neural Network general diagram

The artificial neuron given in this figure has *N* input, denoted as *u1, u2, …, u N*. Each line connecting these inputs to the neuron is assigned a weight, which is denoted as *w1, w2, …, w N* respectively.

The threshold in artificial neuron is usually represented by Φ and the activation is given by this formula:

$$a = \left(\sum_{j=1}^{n} w_j u_j \right) + \Phi$$

The inputs and weight are real values. A negative value for a weight indicates an inhibitory connection while a positive value indicating excitatory one. If Φ is positive, it is usually referred as bias. For its mathematical convenience (+) sign is used in the activation formula. Sometimes, the threshold is combined for simplicity into the summation part by assuming an imaginary input u0 = +1 and a connection weight w0 = Φ. Hence the activation formula becomes:

$$a = \left(\sum_{j=1}^{n} w_j u_j \right)$$

The neuron output function f(a) can be:

Linear: $f(a) = K(a)$

Sigmoid: $f(a) = 1/(1 + \exp(-ka)$

Function implementations can be done by adjusting the weights and the threshold of the neuron. Furthermore, by connecting the outputs of some neurons as inputs to the others, neural network will be established, and any function can be implemented by these networks. The last layer of neurons is called the output layer and the layers between the input and output layer are called the hidden layers. The input layer is made up of special input neurons,

transmitting only the applied external input to their outputs. In a network, if there is only the layer of input nodes and a single layer of neurons constituting the output layer then they are called single layer network. If there are one or more hidden layers, such networks are called multilayer networks.(Partial et al. 2003).

2.4.3 Feedforward neural network

In this kind of networks, the neurons are organized in the form of layers. The neurons in a layer get input from the previous layer and feed their output to the next layer. In this kind of networks connections to the neurons in the same or previous layers are not permitted. Figure 2.6 shows typical feedforward neural network.

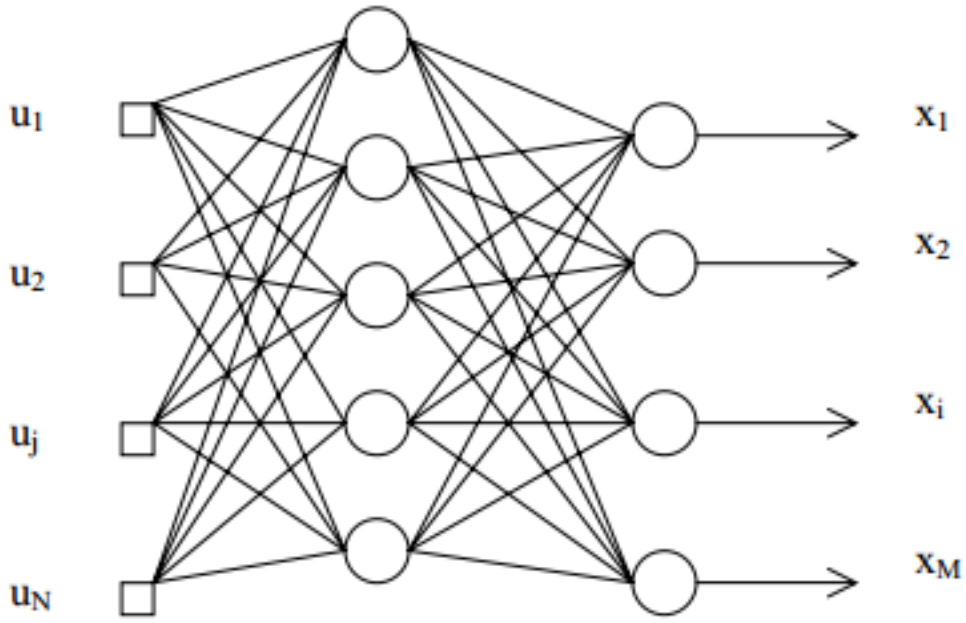


Figure 2.2: Feed forward neural network

For a feedforward network always exists an assignment of indices to neurons resulting in a triangular weight matrix. Furthermore, if the diagonal entries are zero this indicates that there is no self-feedback on the neurons. However, in recurrent networks, due

to feedback, it is not possible to obtain triangular weight matrix with any assignment of the indices.

2.4.4 Training the Neural Network

Neural networks have been trained to perform complex functions in various fields of application including pattern recognition, identification, classification, speech, vision and control systems.

After loading and cleaning the dataset, the feature vectors are calculated for each follower type in the dataset. These feature vectors are used as inputs to train the networks, see Figure 2.3. In training algorithm, the IRIS flower feature vectors that belong to same class are used as input to the network, the target will be binary number for every class in dataset example for class one that have id 1:

Target= [1,0,0], class two 2: target = [0,1,0], and so on.

When the new image is come for recognition, its feature vectors are calculated to gets its new descriptors. These new descriptors are feed as input to the neural network and the network are simulated with these descriptors. The outputs are compared if the output as the same id of IRIS flower class in dataset and error less than threshold then the recognition well done successfully.

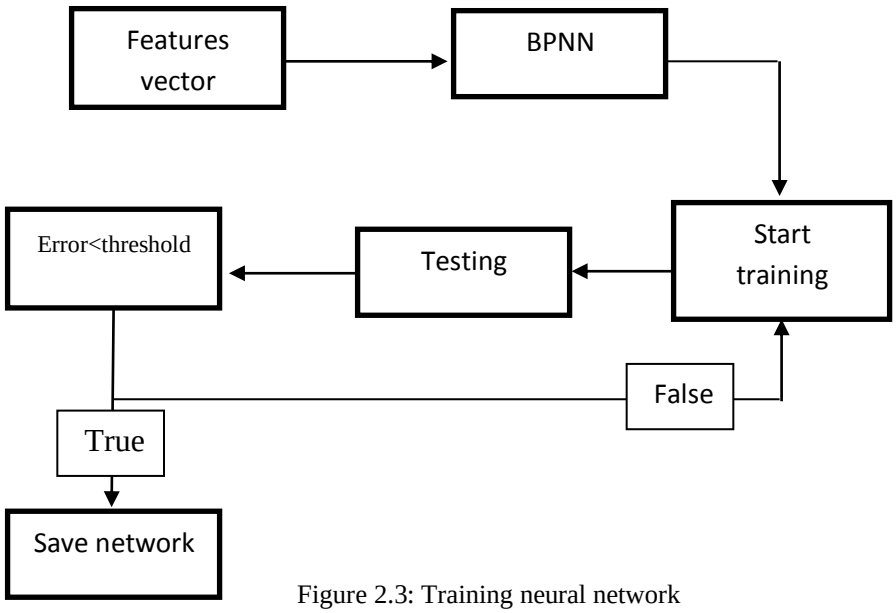


Figure 2.3: Training neural network

2.3 Summary

These steps show the summary of IRIS flower and neural network

1- Load the dataset and clean it to extract the feature vector.
2- Normalize the extracted features.
3- Split the dataset into training set and test set to evaluate the neural network.
4- Create neural network and train it on the training set.
5- For each new IRIS flower image to be identified, calculate its feature vector.
6- Use these feature vectors as network inputs and simulate network with these inputs.

CHAPTER THREE

Source Code

3.1 Programming Language

In our project we use .net environment (C#) and build robust IRIS flowers classification system, the first reason of choice C# because it OOP, huge community, and has rich libraries.

3.2 Graphical User Interface (GUI)

First, we present some of our GUI of the project proposed by the researcher. The Figure 3.1 shows the GUI after loading the dataset with some statics and image thumbnail for each class.

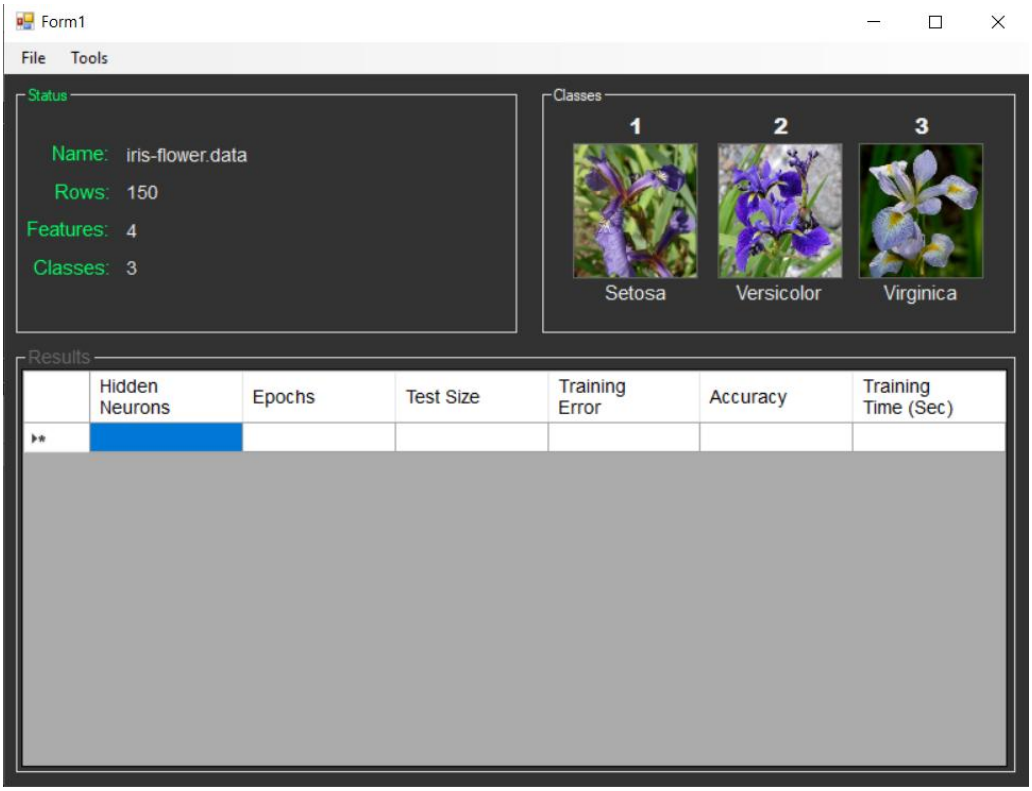


Figure 3.1 First GUI after loading the dataset

The next GUI is the after shuffle the dataset and plot the scatter between all features as shown in the figure 3.2

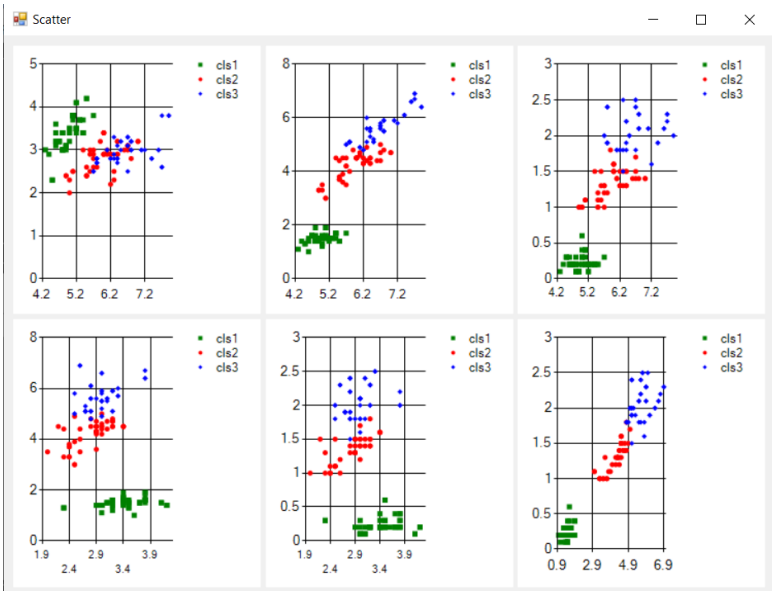
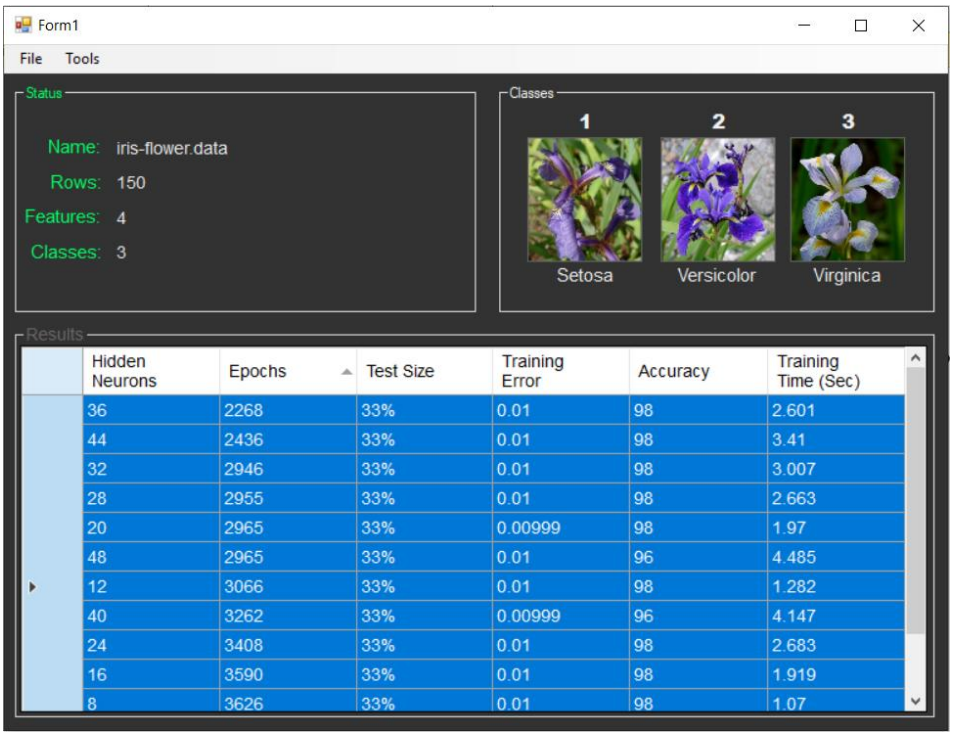


Figure 3.2 Scatter plot for the features

Finally, we the create neural network and train it on the shuffled data and evaluate it on the test set to save the experimental results see Figure 2.3.



Hidden Neurons	Epochs	Test Size	Training Error	Accuracy	Training Time (Sec)
36	2268	33%	0.01	98	2.601
44	2436	33%	0.01	98	3.41
32	2946	33%	0.01	98	3.007
28	2955	33%	0.01	98	2.663
20	2965	33%	0.00999	98	1.97
48	2965	33%	0.01	96	4.485
12	3066	33%	0.01	98	1.282
40	3262	33%	0.00999	96	4.147
24	3408	33%	0.01	98	2.683
16	3590	33%	0.01	98	1.919
8	3626	33%	0.01	98	1.07

Figur3.3 Final results after train and test the network

CHAPTER FOUR

Results and Discussion

The experimental results have been calculated on the trained neural network and the evaluation process depends on the accuracy of the network and the error rate.

4.1 Results

The proposed method is tested on IRIS flower dataset, we use 66.66% of dataset for training process and 33.33% for testing process to ensure our network is working properly.

Our system yields a good result when use IRIS flower features and back propagation neural network, this project has been implemented using C# programming language.

Table 4.1 The experimental results

Hidden Neurons	# Epochs	Test size	Error rate	Accuracy	Training Time
4	**4000**	**33%**	**0.01067**	**98**	**0.691**
8	3626	33%	0.01	98	1.162
12	3066	33%	0.01	98	1.305
16	3590	33%	0.01	98	1.955
20	2965	33%	0.00999	98	1.955
24	3408	33%	0.01	98	2.648
28	2955	33%	0.01	98	2.636
32	2946	33%	0.01	98	2.972
36	2268	33%	0.01	98	2.587
40	3262	33%	0.00999	96	4.105
44	2436	33%	0.01	98	3.377
48	2965	33%	0.01	96	4.392

In the Table 4.1 we notice the top result yield when choose 4 hidden neurons and learning rate was 0.3 and the training time was 0.691 milliseconds. For all training test the

time was performance and speed as well as between 0.6 millisecond to 4.3 milliseconds that show how the network is speed and accurate.

4.2 Discussion

Now we discuss the result of table 4.1

When we increase the number of hidden neurons the training time also increased from 0.6 to 4.3. so, the network hidden neurons must no exceeded the number of input features.

The learning rate usage: if we decrease leaning rate the time of training will be long and if we increase the leaning rate the time will be shorter.

The epochs are decreased when the number of hidden neurons increased, but that affect the training time. Basically, the time is the most important thing beside to the accuracy.

CHAPTER FIVE

Conclusion (and Future work)

In the previous chapters, the proposed methods needed to achieve a system for classify IRIS flower type using neural network and all the results obtained from all proposed approaches and methods were listed.

5.1 Conclusion

This chapter will outline the most critical conclusions reached after studying the problem and the proposed solution, as summarized in the following point

1- IRIS flowers dataset is obtained and pre-processed to extract the features.
2- The feature vector is constructed and shuffled to prepare it for training process.
3- Normalize and split the feature vector into training and testing sets as 33.33% for testing and 66.66% for training.
4- Create neural network with back-propagation algorithm with different hidden neurons size to yield the optimal error rate.
5- Train the network using the training set until the error reach the required error-threshold.
6- Save the neural network with the adjusted weight for later identification.
7- Write the experimental results and discuss it.
8- Finally, suggest a future work.

5.2 Future work

In our project we build an IRIS flowers classification system using IRIS-dataset and the system was implement the identification phase only without detection the flower itself to get result and check our system performance. In the future work we will construct an IRIS detection phase and identification phase using more images and other datasets. Also, in future work we will work in real time video for detection and classification.

References

[1] Aqel, M.M., Jena, R.K., Mahanti, P.K. and Srivastava (2009) 'Soft Computing Methodologies in Bioinformatics', European Journal of Scientific Research, vol.26, no 2, pp.189-203.

[2] Avcı Mutlu, Tülay Yıldırım(2003) 'Microcontroller based neural network realization and IRIS plant classifier application', International XII. Turkish Symposium on Artificial Intelligence and Neural Network

[3] Cho, Sung-Bae.and Dehuri, Satchidananda (2009) 'A comprehensive survey on functional link neural network and an adaptive PSO–BP learning for CFLNN, Neural Comput & Applic' DOI 10.1007/s00521-009-0288-5.

[4] Fisher, A. W., Fujimoto, R. J. and Smithson, R. C.A. (1991) 'A Programmable Analog Neural Network Processor', IEEE Transactions on Neural Networks, Vol. 2, No. 2, pp. 222-229.

[5] Fu, L.(1991) ' Rule learning by searching on adapted nets. In Proceedings of National Conference on Artificial Intelligence' Anaheim, CA, USA, pp. 590-595.

[6] Han, J. and Kamber, M. (2000)' Data Mining: Concepts and Techniques' , 2nd ed. Morgan Kaufmann.

[7] Dr.Hapudeniya, Muditha M. MBBS(2010),'Artificial Neural Networks in Bioinformatics' Medical Officer and Postgraduate Trainee in Biomedical Informatics , Postgraduate Institute of Medicine , University of Colombo ,Sri Lanka, vol.1, no 2,pp.104-111.

[8] Kavitha Kannan 'Data Mining Report on IRIS and Australian Credit Card Dataset', School of Computer Science and Information Technology, University Putra Malaysia, Serdang, Selangor, Malaysia.

[9] Marček D., 'Forecasting of economic quantities using fuzzy autoregressive models and fuzzy neural networks', Vol.1, pp.147-155.1

[10] Pai, G. V and Rajasekaran, S, (2006), 'Neural Networks, Fuzzy Logic and Genetic Algorithms Synthesis and Applications', 6th ed, Prentice Hall of India Pvt. Ltd.

[11] Rath, Santanu and Vipsita, Swati (2010) 'An Evolutionary Approach for Protein Classification Using Feature Extraction by Artificial Neural Network', Int'l Conf. on Computer & Communication TechnologyıICCCT'10.

[12] Towell, G.G. and Shavlik, J.W. (1993), 'Extracting refined rules from knowledge-based neural networks' Mach. Learn, Vol.13, pp.71-101.

[13]Towell, G.G; Shavlik, J.W. (1994) 'Knowledge-based artificial neural networks', Artif. Intell. Vol.7, pp.119-165

الملخص

التصنيف هو تقنية تعلُم الآلة تُستخدم للتنبؤ بعضوية المجموعة لحالات البيانات. لتبسيط مشكلة تصنيف الشبكات العصبية يتم إدخالها. يركز هذا التقرير على تصنيف نبات IRIS باستخدام الشبكة العصبية. تتعلق المشكلة بتحديد أنواع نباتات IRIS على أساس قياسات خصائص النباتات. سيكون تصنيف مجموعة بيانات IRIS هو اكتشاف الأنماط من فحص حجم البتلة والحجر من مصنع IRIS وكيف تم التنبؤ من تحليل النمط لتشكيل فئة .IRIS باستخدام هذا النمط والتصنيف ، يمكن التنبؤ بالبيانات غير المعروفة بشكل أكثر دقة في السنوات القادمة. تم تطبيق الشبكات العصبية الاصطناعية بنجاح على المشاكل في تصنيف الأنماط وتقريب الوظائف والتحسين والذكريات الترابطية. في هذا العمل ، يتم تدريب شبكات إعادة توجيه التغذية متعددة الطبقات باستخدام خوارزمية تعلم الانتشار الخلفي. أظهرت النتائج التجريبية أن معدل الخطأ الأدنى كان 0.01067 مع وقت تدريب 0.691 مللي ثانية ، وكان عدد الخلايا العصبية المخفية 4.

جمهورية العراق
وزارة التعليم العالي و البحث العلمي
جامعة بغداد - كلية العلوم
قسم علوم الحاسوب

تصنيف زهور القزحية باستخدام الشبكة العصبية

مشروع

من قبل
م.م سرى عبد سراب

شعبان 1441

2020 June