Mining Meteorological Data for Prediction of Pakistan Summer Monsoon Rainfall

Submitted by

Jehangir Ashraf Awan

(253-FAS/MSCS/F05)

Supervisor

Dr. Onaiza Maqbool

Assistant Professor
Quaid-e-Azam University, Islamabad

Co-Supervisor

Muhammad Imran Saeed

Assistant Professor
International Islamic University, Islamabad

Department of Computer Science
Faculty of Basic and Applied Sciences
International Islamic University, Islamabad

(2009)

IN THE NAME OF

ALMIGHTY ALLAH

THE MOST BENEFICENT

THE MOST MERCIFUL

Department of Computer Science
International Islamic University Islamabad

24th August, 2009

FINAL APPROVAL

It is certified that we have read the thesis submitted by Mr. Jehangir Ashraf Awan Reg. No. 253-FAS/MSCS/F05 and it is our judgment that this thesis is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad for the MS Degree in Computer Science.

COMMITTEE

External Examiner

Prof. Dr. Nazir Ahmed Sangi
Chairman Department of Computer Science
Allama Iqbal Open University, Islamabad.

Internal Examiners

Asim Munir
Assistant Professor
International Islamic University, Islamabad.

Supervisor

Dr. Onaiza Maqbool
Assistant Professor
Quaid-e-Azam University, Islamabad.

Onaiza Maqbool

Co-Supervisor

Muhammad Imran Saeed
Assistant Professor
International Islamic University, Islamabad.

A dissertation submitted to the
Department of Computer Science,
Faculty of Basic and Applied Sciences,
International Islamic University, Islamabad, Pakistan,
as a partial fulfillment of the requirements for the award of the degree of

MS in Computer Science

DEDICATED

TO

MY LOVING PARENTS

DECLARATION

I hereby declare that this thesis, neither as a whole nor as a part thereof has been copied from any source. It is further declared that I have developed this thesis entirely on the basis of my personal efforts made under the sincere guidance of my supervisors. No portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning.

Jehangir Ashraf Awan
(253-FAS/MSCS/F05)

Acknowledgment

All praise to Almighty Allah, the most merciful, the most beneficent, Who blessed me with opportunity, health and courage to learn and do this work.

My deepest gratitude and appreciation goes to Dr. Onaiza Maqbool, Asst. Professor Quaid-e-Azam University for her guidance, advice, timely motivation and creative suggestions to carry out and complete this study.

My sincere appreciation goes to Asst. Professor Imran Saeed for his suggestions and helpful assistance to complete this work.

I am also indebted to Dr. Ghulam Rasul for his valuable suggestions, guidance and constructive criticism. I would like to thank all my colleagues at Pakistan Meteorological Department specially Afzaal Karori and Khurrum Waqas Haider for their help and encouragement.

I would like to extend my deepest gratitude to my loving parents for their prays, encouragement and support in my studies and whole of my life without that I would have not been able to accomplish any thing worthwhile. My brothers Zahir Akhtar Awan and Shakeel Ashraf Awan deserve special mention and my sincere thanks.

I am grateful to my father in law for his constant encouragement and valuable suggestions to complete this work.

Finally I pay my special thanks to my caring wife and daughter for their support and encouragement.

Abstract

Data mining is an emerging field that is becoming popular for discovering knowledge in many domains. However its use for Meteorological data has been limited. This study makes use of data mining for predicting Pakistan Summer Monsoon Rainfall (PSMR). A significant amount of Pakistan's annual rainfall comes from summer monsoon. Monsoon has great impact on agriculture, which is the backbone of our country's economy. Extreme monsoon years are the major cause of floods and droughts in Pakistan.

Presently, regression based statistical models and downscaling models i.e. statistical downscaling and dynamical downscaling are commonly used for monsoon rainfall prediction. Neither of these techniques has been satisfactory in predicting monsoon rainfall. This study investigates use of two neural network algorithms i.e. Backpropagation (BP) and Learning Vector Quantization (LVQ) for PSMR prediction. Suitable BP and LVQ models are obtained by applying several variations of window size and hidden/competitive layer neurons. Forty three years monsoon total rainfall (July, August & September) data from 1960 to 2003 is used for training of neural network models to predict PSMR from 2004 to 2008.

The viability of proposed models is demonstrated by comparing the predicted results with actual monsoon rainfall as well as with the results of Pakistan Meteorological Department (PMD)'s statistical model and statistical downscaling technique. This comparison reveals the better performance of proposed neural network models as compared to traditional statistical techniques, in terms of accuracy as well as in terms of resources needed to predict monsoon rainfall.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1. Introduction

Rainfall is a versatile phenomenon which is the last product in a series of delicate and non-linear phenomena. Monsoon is the major source of rainfall in Pakistan. The term Monsoon is derived from the Arabic word "Mausam", which means season. Monsoon enters Pakistan in the month of July and lasts till the month of September. About 59 percent of annual rainfall of Pakistan comes from summer monsoon [1].

Monsoon has great affect on agriculture, water resources and human lives. Intense monsoon is the major cause of flooding, while weak monsoon leads to drought. The destruction due to floods in Pakistan has been enormous. According to the report of Federal Flood Commission, Pakistan, more than 7,200 people have lost their lives in fifty nine years from 1950 to 2006 because of floods [2].

Although monsoon may cause devastating affects, it is a blessing when normal, as it is the major source of water, which brings up the level of water reservoirs and caters to the needs of drinking water as well as agriculture.

Correct and timely monsoon rainfall prediction is crucial for devising various policies e.g. for agriculture, as well as for administrative purposes. These policies, when devised and implemented in a timely manner may reduce the destruction caused by monsoons. Commonly used techniques for monsoon rainfall predictions are regression based statistical models and dynamic/statistical downscaling models. Statistical models have been in use for a number of decades to predict monsoon rainfall. Gilbert Walker [3] introduced regression based statistical model for Indian monsoon rainfall prediction. Although India has made several efforts to enhance these statistical models for Indian Monsoon Rainfall prediction [4] - [11], the results of these models have not improved over seven decades [12]. On the other hand, Downscaling models i.e. dynamical downscaling and statistical downscaling, downscale the output of General Circulations Models also known as Global Climate Models (GCM) to regional scale for prediction. GCM are numerical models that have grid resolution of 150km to 300km [13] and require huge processing power to simulate the global atmosphere. Inspite of making use of such a large number of resources and having dependency on a number of parameters, the results of these models for monsoon rainfall have not been satisfactory. The problem is specially acute in extreme years of monsoon i.e. flood or drought [12].

Data mining is an emerging field that has shown promising results in a variety of domains including medicine, security, judiciary, biometrics, multimedia retrieval, scientific data analysis, marketing, land use and geological studies [14], and is specially suited for large datasets. However its application in meteorological data has been limited. There are studies for Indian Monsoon rainfall prediction using Neural Network [15] - [20], and also for Bangladesh monsoon rainfall prediction [21], but there has been no work conducted in this area for Pakistan Summer Monsoon rainfall prediction.

In this dissertation we employ neural network algorithms for Pakistan Summer Monsoon Rainfall prediction. Our contributions are detailed in the next Section.

1.1 Research Contributions

i) Till now only regression based statistical and statistical downscaling techniques have been used for Pakistan Summer Monsoon Rainfall (PSMR) prediction. We investigate two neural network algorithms i.e. Backpropagation (BP) and Learning Vector Quantization (LVQ) for PSMR prediction and evaluate whether they are capable of producing better results than statistical and dynamical techniques.

ii) Appropriate configuration of a neural network for the problem at hand is necessary. We discuss the neural network implementation issues in detail i.e. selection of appropriate window size, number of hidden/competitive layer neurons and neural network generalization for PSMR prediction.

iii) Real data has great importance for evaluating the viability of any model. We used forty nine years (1960-2008) real monsoon (Jul, Aug, Sep) rainfall data for training and testing of the models, and predicted monsoon rainfall of Islamabad and All Pakistan not only for 5 years (2004-2008) test period but also for coming year 2009.

iv) Comparative studies reveal the strengths and weaknesses of models. We compare the results of both neural network techniques i.e. BP and LVQ, not only with actual monsoon rainfall data but also with the results of Pakistan Meteorological Department (PMD)'s Statistical monsoon rainfall prediction model and statistical downscaling model.

1.2 Thesis Outline

This thesis is composed of five chapters. Chapter 2 covers the relevant background and literature survey. Chapter 3 presents the overview of techniques already in use and proposed neural network techniques for PSMR prediction. Chapter 4 presents the experimental setup and results. This Chapter includes the comparison, comprehensive evaluation of techniques and detailed analysis of results. Chapter 5 presents the conclusions of the thesis, and recommendations for future work.

Chapter 2

Literature Review

2. Literature Review

In this chapter we present the work of researchers in the field of data mining and statistics for monsoon rainfall prediction. Review of literature that gives application of data mining in meteorological data for several other purposes i.e. Electricity load forecasting and temperature prediction, is also presented. It should be noted that Indian subcontinent is the mostly affected region due to monsoons. That is the reason that most of the work is carried out in India for monsoon rainfall prediction.
Literature survey is divided into three Sections. Section 2.1 presents the research work that has been carried out for monsoon rainfall prediction using statistical approach. In Section 2.2 we review the work of researchers in the field of neural networks for monsoon rainfall prediction. Section 2.3 illustrates the applications of data mining techniques on meteorological data.

2.1 Statistical Approach to Predict Monsoon Rainfall

Sir Gilbert Walker [3] was the first who introduced regression based statistical technique to predict Indian monsoon rainfall. Later many scientists made efforts to improve forecasting by introducing different set of predictors and datasets [4] - [11].

In spite of these efforts, statistical models, which are being used operationally by Indian Meteorological Department since 1932 have not improved over seven decades [12]. A study by Gadgil *et. al* [12] indicates that the results of both dynamical models and statistical techniques are not satisfactory for monsoon prediction and results are worse for extremes i.e. droughts and excess rainfall seasons.

2.2 Neural Networks to Predict Monsoon Rainfall

In 2008, Guhathakurta used artificial neural networks (ANN) for India sub-divisions monsoon rainfall prediction [15]. He used separate three-layer backpropagation neural networks with one hidden layer having three neurons, for each of the 36 homogeneous meteorological sub-divisions of India. Guhathakurta used data from 1941-1991 for training of model and 12^{th} year India summer monsoon rainfall was predicted using past 11 year monsoon rainfall data. All India summer monsoon rainfall was predicted using separate neural network model. ANN model was also developed for average area weighted monsoon rainfall forecast of each of 36 sub-divisions. The performance of these models is encouraging. However results from area weighted model were closer to actual than separate ANN model for whole of India and it also captured the excess and deficient years accurately.

In [16], eight parameters probabilistic ANN model is developed for monsoon rainfall prediction of district Ambikapur Chhattisgarh. Eight regional and global parameters i.e. Mean Vapour Pressure (in hPa), Highest Maximum Temperature (in Deg C), Mean Station Level Pressure (in hPa), Mean Relative Humidity (in %), Mean Total Cloud Amount (in Octas), Mean Wind Speed (in kmph), Mean Wet Bulb Temperature (In Deg. C) and Mean Maximum Temperature(in Deg. C) are used as model input. The ANN model is constructed using 11 input neurons, 11 hidden layer neurons and 01 output neuron. Results of this study shows that eight parameters probabilistic model predicted monsoon rainfall well and has error less than or equal to one third of standard deviation. This study reveals that neural network technique has the capability to identify internal dynamics of rainfall successfully.

In [17], three-layer feedforward backpropagation neural network algorithm is used to predict monsoon rainfall of Eastern Plateau Mahanadi Basin-EPMB Chhattisgarh subdivision. This study used monsoon rainfall data from 1945 to 1995 for training and predicted monsoon rainfall from 1996 to 2006 using deterministic and probabilistic approach. In probabilistic approach six predictors i.e. Mean Vapour Pressure (in hPa), Mean Station Level Pressure (in hPa), Mean Relative Humidity (in %), Mean Sea Level Pressure (in hPa), Mean Wet Bulb Temperature (in Deg C) and Mean Maximum Temperature (in Deg. C) were used to predict monsoon rainfall. Results of deterministic model gave error less than or equal to half of the standard deviation, with correlation above 0.8 for the districts and 0.7 for the whole subdivision. Results of probabilistic model were slightly better than those of deterministic model.

In [18] ANN is used to predict the monsoon rainfall for 14 districts of subdivision Kerala. This study used observed data of 36 stations from 1941-1991 for model training and from 1992 to 2004 for testing of the model. Three layer backpropagation Neural Network was used with three hidden layer neurons and one output neuron. Past eleven years data was used to predict the 12th year monsoon rainfall. Forecast for subdivision Kerala was made both by using separate ANN model and using area weighted values of all districts forecast. The RMSE obtained for predicting monsoon was less than standard deviation for all districts and even less than one-third for some districts. Results of this study show that neural network performed well for smaller spatial scale and has longer lead time of predicting monsoon almost a year in advance.

Other researchers who have employed neural networks for Indian monsoon rainfall prediction include [19], [20].

In [21], ANN, Adaptive Neuro Fuzzy Inference System (ANFIS) and Genetic Algorithm (GA) are used for Bangladesh Monsoon rainfall prediction. Prediction is made for the stations of Barishal, Chittagong, Dhaka, Khulna, Rajshahi and Sylhet. Parameters used as predictors for monsoon prediction are temperature, relative humidity and cloud coverage for monsoon period (April to October). Results of this study show that ANFIS performed better than other proposed models.

2.3 Data Mining and Meteorological Data

Apart from monsoon prediction, Data Mining techniques have been utilized for mining meteorological data for several other purposes all over the world.

McCullagh *et. al* [22] investigated ANN by dividing rainfall estimation problem into different expert networks. He developed separate expert network for each rainfall band i.e. high, low and medium rain. Results from these separate expert networks show 10.44%, 18.95% and 11.55% classification improvement over standard backpropagation network in low, medium and high rainfall bands respectively.

Kotsiantis *et. al* [23] used different data mining techniques for estimation of minimum, maximum and average daily temperature. They investigated different algorithms i.e. Linear Regression, Model trees (M5'), M5rules, Instance based learners (IB3), Backpropagation (BP) and Additive regression with training datasets of previous one, two and three years, for prediction of year 2005. They found that regression algorithms gave satisfactory results using input of previous years' temperature values. Their results also indicate that two years historical temperature data is enough to predict minimum, maximum and average temperature, and there is no need of using larger training datasets.

Dong-Xiao *et. al* [24] used Vary Structure Neural Network (VSNN) for electricity load forecasting using historical weather data and electricity load. This approach determines weather characteristics of load forecasting day using weather forecast and then maps it to historical load days having same weather characteristics using different data mining techniques. Finally a three layer VSNN model for electricity load forecasting is developed. This approach shows improved accuracy in daily load forecasting as compared to linear regression and common ANN model.

In [25] ANN is used to forecast load of Liaoning power system. In this study historical load data is divided according to weather characteristics and forecast was made using load information of historical load days having weather characteristics similar to forecasting day. Gray relevant analysis was used to extract historical daily load data having same characteristics as forecasting day. Number of layers for BP network was selected according to Kolmogorol theorem. This study shows that pretreatment for the historical data reduces the training time for BP network and proper adjustment of the number of hidden layers and optimal network model increases accuracy of forecasting. Results obtained using ANN was satisfactory for Liaoning power system.

In [26] a grid model in terms of geographical division is presented for short-term electricity load forecasting. Grid model of short-term electrical load forecasting was established by dividing a big network of North China into small subnets Beijing, Tianjin and Tangshan. In this study three-hourly data was used from June 2003 to May 2004 to establish load forecast model and from July 2004 to December 2004 for adjustment of established model. Each subnet model was developed using mining default rules based on rough set (MDRBR) method. MDRBR algorithm was used to apply mining policy from bottom to top. Multi-layered network of rules was designed having one conditional attribute in first layer, two conditional attributes in second layer and so on until specified threshold achieved or all attributes done. Attributes selected as conditional attributes were holiday, weekly type, temperature, rate of temperature change, humidity, rate of humidity change, previous day load, previous two day load, previous three day load and previous week load. Temperature was divided in high, middle & low and rainfall was divided to be zero or non-zero categories. This study used equalizing value and variance to identify abnormal data, and conditional-mean to modify abnormal data and fill the missing value. Using grid model the efficiency of rules on average was 89.92% while using the non-grid model it was only 76.74% and accuracy rate of load forecasting using grid model was 98.01% while the non-grid was only 91.68. Results of this study reveal that proposed grid model is much better than non-grid model.

In [27] Sequential Minimal Optimization (SMO) algorithm is used to forecast short term power load. Weather parameters used in this study were high temperature, lowest temperature, average temperature and daily rainfall. Weather data was divided into four categories i.e fine, cloudy, rain and snow, and load data was divided into three categories i.e workday, weekend and important holiday. This approach used load information of historical load days having same weather characteristics as forecasting day. Forecast of power load was made for seventeen days from 10-07-2005 to 26-07-05 using SMO algorithm. Forecasted results reveal that the SMO is more accurate than BP neural network.

Many researchers [28] - [30] have made similar attempts to forecast electricity load using weather data and different Data Mining techniques.

In [31] K-Nearest Neighbour (KNN) is used to classify historical weather data and predict the weather. In this study historical weather data for ten years was used that includes several weather parameters i.e. temperature, Maximum Temperature, Minimum Temperature, Wind Speed, Maximum Wind Speed, Minimum Wind Speed, Dew Point, Sea Level Pressure, Snow Depth and Fog. To clean noisy and missing data, Means and Bin Means data mining techniques were used and textual values i.e. Yes/No were replaced with number values i.e. 1/0. Prediction was made for 20 days from 15-02-2000 to 05-03-2000 using different values of K. The prediction accuracy was 96.66% for attributes having Boolean values i.e. Fog, Hail, snow, ice, Thunder. This study demonstrates that The KNN can predict up to seventeen climatic attributes, i.e. Mean Temperature, Max Temperature, Min Temperature, Sea Surface Temperature, Sea Level Pressure, Gust etc. at the same time and none of the previous developed systems can predict such a huge set of attributes at the same time with such level of accuracy.

Chapter 3

Overview of Existing and Proposed Techniques

3. Overview of Existing and Proposed Techniques

Currently both dynamic and statistics techniques are being used in a race to accurately predict the weather. Dynamic prediction involves large investment in computer hardware as well as costly appliances related to it. To establish such a unit, which also needs constant up-gradation to cope with increasing computation demand, is a difficult, costly and time consuming procedure. On the other hand, statistical techniques are far more economical. What we need in this regard is a historical dataset which is easily available with National Meteorological Services Department.

Although statistical techniques are more economical to use as compared to dynamical techniques, neither technique has been successful in predicting monsoon rainfall accurately because of its high spatio-temporal variability and complex topography of region.

Neural networks have shown promising results in variety of domains. Neural networks have ability to tolerate noisy data and can predict and classify patterns that they have not been trained on [32].

In this chapter, we briefly explain the proposed neural network techniques i.e. Backpropagation and LVQ. We also give an overview of techniques that are currently being used for Pakistan Summer Monsoon Rainfall prediction i.e. Statistical Downscaling Technique and PMD's statistical multiple regression based model.

This chapter is organized as follows. In Section 3.1, statistical downscaling technique is elaborated. In Section 3.2, statistical model of PMD for monsoon rainfall prediction is presented. Section 3.3 defines Data Mining. Neural Network details are given in Section 3.4. Finally, Section 3.5 presents the model fitness measures.

3.1 Statistical Downscaling Technique

General Circulation Models (GCM) are mathematical models that predict global atmosphere and/or oceanic variability. These models require huge processing power to simulate the atmosphere. GCM has a coarse grid resolution of 150km to 300km [13] and has global spatial coverage. To predict local or regional climate using GCM output, the method used is downscaling. Downscaling can be accomplished in two ways, first using dynamical downscaling and second using statistical downscaling [33]. Dynamical Downscaling is done by nesting the finer resolution regional model within GCM. It requires huge processing power and time as in the case of numerical models. On the other hand, Statistical Downscaling technique deduce regional climate by developing relationships (e.g. multiple regression) between GCM output and local observational data.

The Pakistan Meteorological Department has carried out experiments using statistical downscaling for seasonal prediction of rainfall over different stations. Multiple linear regression model is developed for this purpose. This model takes its inputs from GCM which then are statistically downscaled using multiple linear regressions. The global coupled ocean-atmosphere model of National Climate Centre, CMA, China, is being used. Seasonal accumulative rainfall is being used as predictand while different fields of GCM output are used as predictor [13].

This technique is dependent on output of GCM. GCM in turn requires huge processing power and time to perform model simulations. These techniques are based on the assumptions that global atmosphere is being well predicted by GCMs, thus error in GCM would further the error of this technique.

3.2 Statistical Technique

PMD started Pakistan Monsoon Rainfall prediction in 1950 [1]. This prediction is based on linear regression formula introduced by G.W Walker [3].

The long-range monsoon rainfall prediction model of PMD uses the following multiple regression equation.

$$R \quad = \quad 0.3575 * A - 14.3989 * B - 0.7612 * C - 0.8147 * D + 0.3902$$

Where

R = Pakistan monsoon rainfall departure from the normal monsoon rainfall.

A = South American Pressure departure.

Mean of the station level pressure of Buenos Aires, Cordoba and Santiago for the months of April and May.

B = Equatorial Pressure

(i) Mean monthly station level pressure of Secheylles for March, April and May.

(ii) Mean sea level pressure of Port Darwin for February and March.

C = Mean range of temperature at Lahore, Islamabad, Sialkot, Multan for the months of April and May.

D = Western Himalayas Snow accumulation in the month of May.

Similar to downscaling technique, statistical model is dependent on number of variables. That limits its capability of predicting monsoon beyond one month lead time.

3.3 Data Mining

The following material to elaborate this technique is taken from [34] - [36]. For further details, the reader may refer to these sources.

Data volumes have boosted with the advancements in computing and data acquisition technology. This growth is evident in many domains e.g. Meteorological Data, Satellite imagery, medical records, phone call details, credit card usage and supermarket transactions. The rapid growth in data volumes has raised the interest of datasets holders in seeking useful information from these huge datasets. Dealing with such huge datasets is a challenging issue. It is not trivial task to extract useful information/knowledge from gigabytes to terabytes of data. Data mining is an emerging field that gained the interest of users to achieve this task.

Data mining is the process of extracting useful, hidden patterns from large amounts of data [32]. Data mining tasks can either be predictive or descriptive. Predictive data mining deals with predicting values of data using known results found from data. It may base on use of historical data. Predictive data mining tasks include classification, regression, time series analysis and prediction. On the other hand descriptive data mining describes the patterns and relationships in data. Its goal is to explore the features of data, not to predict new values. Descriptive data mining tasks include clustering, summarization, association rules and sequence discovery.

3.4 Artificial Neural Network

The following material to elaborate this technique is taken from [37] - [38]. For further details, the reader may refer to these sources.

Neural networks are the most widely known data mining technique [39]. They gained popularity because of their highly accurate predictive models [40]. Artificial neural networks (ANNs) are based on biological neural networks. A Neural network resembles the brain in two respects, first it acquires knowledge from its environment through a learning process, and second weights on interneuron connections are used to store the acquired knowledge [38]. Neural Network comprises of processing units called neurons, communication links to interconnect neurons and weights associated with each communication link. A simple neural network is illustrated in Figure 3.1.

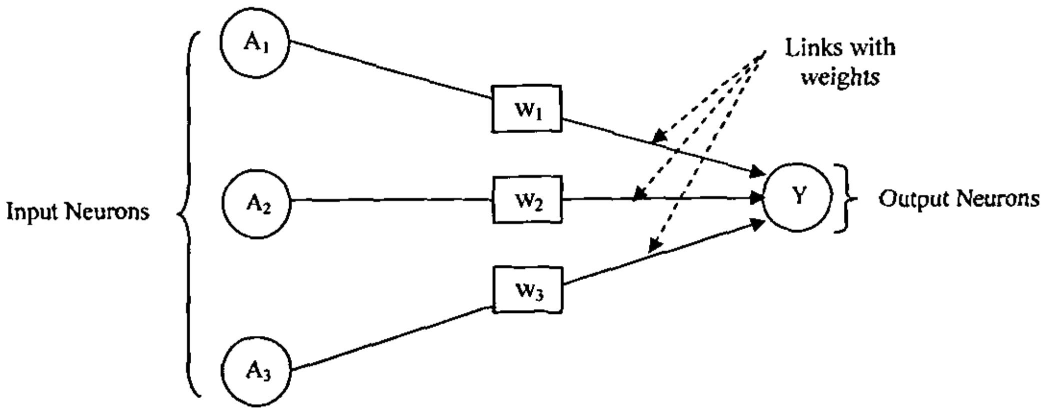

Figure 3.1: Simple Neural network having 03 input neurons and one output neuron interconnected using weighted links.

Each neuron has an activation state (that is the function of its net input), which determines output of that neuron. Activation functions are discussed in detail in Section 3.4.5. In Figure 3.1 neuron "Y" takes input from Input neurons A_1, A_2, A_3. Output of neurons A_1, A_2, A_3 is a_1, a_2, a_3 and weights are w_1, w_2, w_3 respectively. The net input to neuron "Y" is given by

$Y\_input = a_1w_1 + a_2w_2 + a_3w_3$

The output of neuron Y by applying binary step activation function is

$$Y = \begin{cases} 1 & \text{if } f(\,y\_input\,) > 0 \\ 0 & \text{if } f(\,y\_input\,) \leq 0 \end{cases}$$

3.4.1 Neural Network Architecture

The composition of neurons in layers i.e. input layer, hidden layer and output layer, and connection patterns within and between layers is called neural network architecture. Neural networks are often divided into two categories, single layer or multilayer. While counting the number of layers, input layer is not counted as a layer, because it does not perform any processing.

3.4.1.1 Single and Multilayer Neural Network

Neural network having one input and one output layer is called single layer network. An example of single layer network is presented in Figure 3.2.

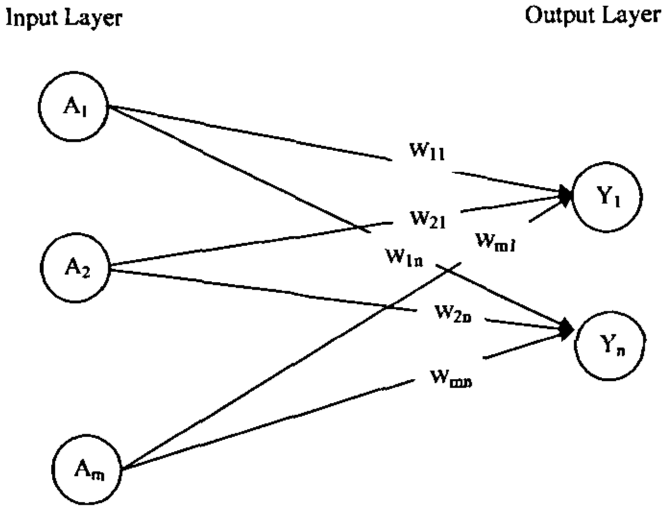

Figure 3.2: Single Layer Neural Network

For complex problems multilayer neural networks are required. Multilayer neural networks consist of one or more hidden layers in addition to an input layer and an output layer. Hidden layers comprising of hidden neurons lie in between input layer and output layer. An example of multilayer neural network is presented in Figure 3.3.

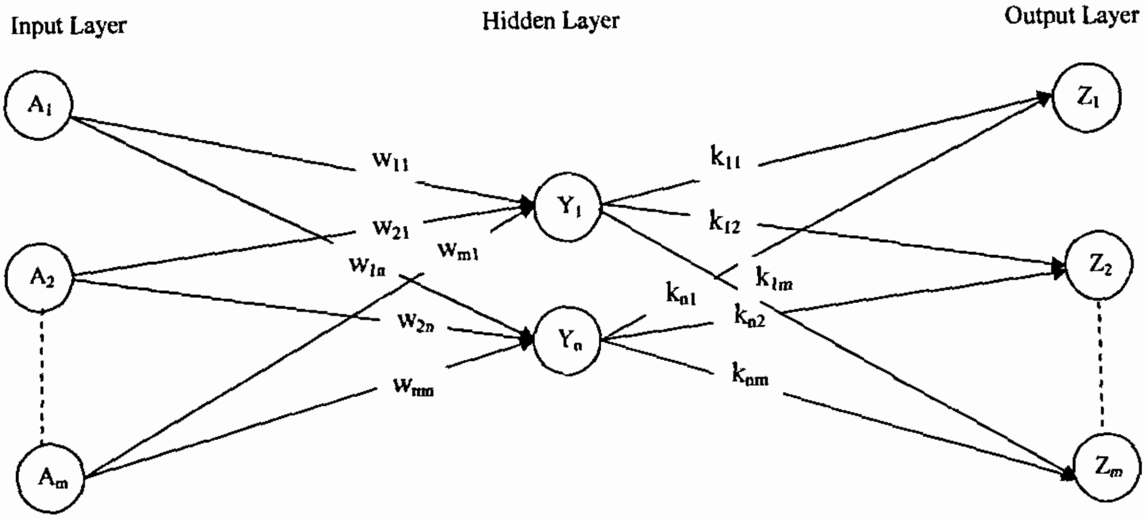

Figure 3.3: Multilayer Neural Network

3.4.2 Neural Network Training and Prediction

Similar to biological neural network, ANN requires training to operate. Neural network training is the procedure of adjusting weights using a dataset called training dataset.

Training is often carried out by presenting sequence of input vectors with target output vectors to the network. This kind of training is known as supervised training. Training in which target vector is unknown is said to be unsupervised training.

Network, once trained, is presented with unseen dataset known as test dataset for prediction or classification tasks. Trained Network uses adjusted weights to predict target vectors of unseen input vectors.

To perform the task of training and prediction specific algorithms are required. Neural network training algorithms used in this study are Backpropagation and Learning Vector Quantization. These algorithms are described in Section 3.4.7. Before describing these training algorithms, it is required to define certain terms i.e. Learning Rate, Bias, Momentum Coefficient and Activation Functions.

3.4.3 Learning Rate

Learning rate α controls the rate of weight change during training of neural network. It is adjusted to small random values i.e. $0 < \alpha \leq 1$. Very small learning rate slows down learning, while very large learning rate results in oscillations [41].

3.4.4 Bias

A bias can be added like any other weight to the layer. It has constant activation of "1". A bias denoted by "w_0" is presented in Figure 3.4.

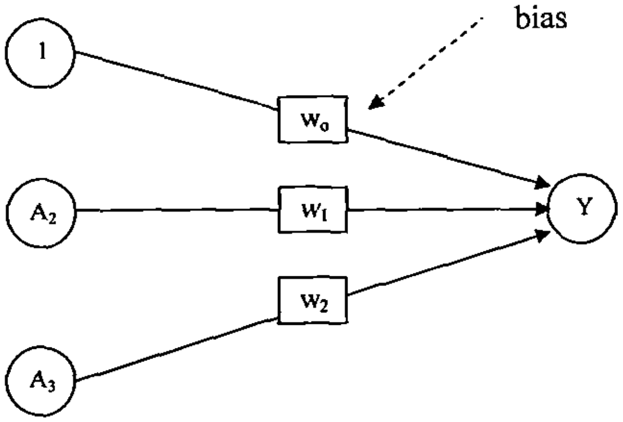

Figure 3.4: A simple neural network with bias w_0

3.4.5 Activation Function

Activation function of neuron, also known as output function, is applied on weighted sum of neuron's net input. For input layer the activation function is identity function. This means that it performs no processing and passes on input signal as it is to next layers. The output layer and hidden layers may use same or different activation functions. Most common activation functions are binary step function and sigmoid function [37].

3.4.5.1 Binary Step Function

Binary step function is used where desired output is either on or off i.e. (0 or 1) or (-1 or 1). Binary step function is also known as threshold function. It converts continuous net input to discrete (0 or 1) or (-1 or 1) using specified threshold θ.

$$Y = \begin{cases} 1 & \text{if } f(x) \geq \theta \\ 0 & \text{if } f(x) < \theta \end{cases}$$

3.4.5.2 Sigmoid Function

Sigmoid Functions are activation functions having "S" shaped curve. One of the most common sigmoid functions is the logistic sigmoid function. Logistic sigmoid function with output in range of (0 to 1) is said to be binary sigmoid given in Figure 3.5, while logistic sigmoid function with output in range of (-1 to 1) is said to be bipolar sigmoid function illustrated in Figure 3.6. These functions are specially suited for backpropagation algorithm because of having simple relationship between value of function and its derivative, which reduces the processing requirements during learning phase.

Binary sigmoid function and its derivative are given below:

$$f(x) = \frac{1}{1+\exp(-\sigma x)}$$

where σ is steepness parameter, that determines the steepness of curve.

$$f'(x) = \sigma f(x)[1 - f(x)]$$

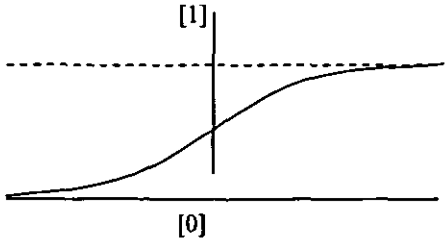

Figure 3.5: Binary Sigmoid

Bipolar sigmoid function and its derivative are given below:

$$f(x) = \frac{1-\exp(-\sigma x)}{1+\exp(-\sigma x)}$$

$$f'(x) = \frac{\sigma}{2}[1 + f(x)][1 - f(x)]$$

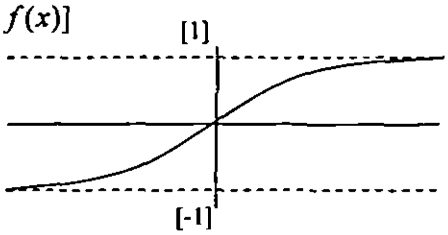

Figure 3.6: Bipolar Sigmoid

3.4.6 Neural Network Training Algorithms

Algorithms used in this study for neural network training are Backpropagtaion and Learning Vector Quantization.

3.4.6.1 Backpropagation

Backpropagation is the most common, neural network algorithm. It consists of a multilayer, feedforward network. Feedforward network is a network in which signals flow in forward direction from input neurons to the output neurons.

Backpropagation learns by iterative processing of training patterns and comparing the network predicted values with actual known targets. Error is calculated and propagated back from output layer to hidden layers, to modify the weights so as to minimize the error between actual targets and predicted values. Backpropagation algorithm has three major steps. Feedforward of input patterns, backpropagation of calculated error and weights adjustment. The Backpropagation training algorithm is presented in Figure 3.7. This algorithm is taken from [37].

Nomenclature used in Backpropagation Algorithm:

x Input training vector: $x = (x_1, \ldots, x_i, \ldots, x_n)$.

t Output target vector: $t = (t_1, \ldots, t_k, \ldots, t_m)$.

δ_k Portion of error correction weight adjustment for w_{jk} that is due to an error at output unit Y_k: also, the information about the error at unit that is propagated back to the hidden units that feed into unit Y_k.

δ_j Portion of error correction weight adjustment for v_{ij} that is due to the backpropagation of error information from the output layer to hidden unit Z_j.

α Learning rate

X_i Input unit i: For an input unit, the input signal and output signal are the same, namely x_i

v_{0j} Bias on hidden unit j.

Z_j Hidden unit j: The net input to Z_j is denoted z_in_j: $z\_in_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$

The output signal (activation) of Zj is denoted zj: $z_j = f(z\_in_j)$

W_{0k} Bias on output unit k.

Y_k Output unit k: The net input to Y_k is denoted y_in_k: $y\_in_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$

The output signal (activation) of Y_k is denoted y_k: $y_k = f(y\_in_k)$

Stopping Condition: The condition may specify a fixed number of iterations or reducing of error to some specified threshold.

Backpropagation Algorithm:

```
Step 0: Initialize weights.
        (Set to small random values)
Step 1: While stopping condition is not satisfied do steps 2 – 9.
        Step 2: For each training pair (input, target) do steps 3 – 8.
        Feedforward:
        Step 3: Each input unit ($X_i$, i = 1, ..., n) receives input signal $x_i$ and broadcasts this
        signal to all units in the layer above (the hidden units)
        Step 4: Each hidden unit ($Z_j$, j=1, ..., p) sums its weighted input signals.
        $z\_in_j = v_{0j} + \sum_{i=1}^{n} x_i v_{ij}$
        applies its activation function to compute its output signal.
        $z_j = f(z\_in_j)$
        and sends this signal to all units in the layer above (output units).
        Step 5: Each output unit ($Y_k$, k = 1, ... , m) sums its weighted input signals.
        $y\_in_k = w_{0k} + \sum_{j=1}^{p} z_j w_{jk}$
        and applies its activation function to compute its output signal.        $y_k = f(y\_in_k)$
Backpropagation of Error:
Step 6: Each output unit ($Y_k$, k = 1 , ..., m) receives a target pattern corresponding to the
        input training pattern, computes its error information term,
        $\delta_k = (t_k - y_k) f' (y\_in_k)$
        calculates its weight correction term (used to update $w_{jk}$ later).
        $\Delta w_{jk} = \alpha \delta_k z_j$
        calculate its bias correction term (used to update $w_{0k}$ later).
        $\Delta w_{0k} = \alpha \delta_k$
        and sends $\delta_k$ to units in the layer below.
Step 7: Each hidden unit ($Z_j$, j = 1 ,..., p) sums its delta inputs (from units in the layer above).
        $\delta\_in_j = \sum_{k=1}^{m} \delta_k w_{jk}$
        multiplies this by the derivative of its activation function to calculate its error
        information term.
        $\delta_j = \delta\_in_j f' (z\_in_j)$
        calculates its weight correction term (used to update $v_{ij}$ later).
        $\Delta w_{ij} = \alpha \delta_i x_i$
        and calculates its bias correction term (used to update $v_{0j}$ later).
        $\Delta w_{0j} = \alpha \delta_j$
        Update weights and biases:
        Step 8: Each output unit ($Y_k$, k = 1 ,..., m) updates its bias and weights (j = 0 , ..., p);
        $w_{jk} (new) = w_{jk} (old) + \Delta w_{jk}$
        Each hidden unit (Zj, j = 1 ,..., p) updates its bias and weights (i = 0 ,..., n);
        $v_{ij} (new) = v_{ij}(old) + \Delta v_{ij}$
Step 9: Test stopping condition.
```

Figure 3.7: Backpropagation Training Algorithm

3.4.6.2 Learning Vector Quantization (LVQ)

LVQ is a competitive supervised neural network for pattern classification. It has competitive layer, in which each neuron competes to win. The winning neuron has output 1 while all other neurons have output 0. The weight vector of an output unit is often called codebook or reference vector. Each codebook vector represents a region for particular class label. When new input is presented, LVQ finds the closest codebook vector and assigns the class label of that code vector to the input. The training algorithm for LVQ is presented in Figure 3.8. This algorithm is taken from [37].

Nomenclature used in LVQ algorithm:

x Input training vector: $x = (x_1, \ldots, x_i, \ldots, x_n)$.

T correct category or class for the training vector.

w_j weight vector for jth output unit $(w_{1j}, \ldots, w_{ij}, \ldots, w_{nj})$.

C_j category or class represented by jth output unit.

$||x\text{-}w_j||$ Euclidean distance between input vector and (weight vector for) jth output unit.

Algorithm

```
Step 0: Initialize reference or codebook vectors
        Initialize learning rate $\alpha$ (0)
Step 1: While stopping condition is false, do steps 2 – 6.
        Step 2: For each training input vector x, do steps 3 – 4.
                Step 3: Find J so that $||x - w_j||$ is a minimum.
                Step 4: Update $w_j$ as follows:
                if $T = C_j$ then
                $w_j$ (new) = $w_j$ (old) + $\alpha[x - w_j(\text{old})]$
                if $T \neq C_j$ then
                $w_j$ (new) = $w_j$ (old) – $\alpha[x - w_j(\text{old})]$
Step 5: Reduce learning rate
Step 6: Test stopping condition:
        The condition may specify a fixed number of iterations (i.e. executions of Step 1) or
        the learning rate reaching a sufficiently small value.
```

Figure 3.8: Training Algorithm for LVQ

3.4.7 Neural Network Configuration issues and its Input

Before training neural network using training algorithms, neural network is required to be configured properly. This includes: (i) neural network generalization and preprocessing of network input (i.e. normalization and sliding window approach). (ii) defining number of hidden layers neurons. These are important issues to be discussed for improving neural network performance.

3.4.7.1 Normalization

Data Normalization is an important preprocessing phase. It helps in speeding up neural network learning [32]. Normalization is done by scaling a value to some specific range, typically [0 to 1] or [-1 to 1].

There are two most common data normalization methods.

a) Min and Max

b) Mean and Standard Deviation

a) Min and Max

Data is scaled to range from -1 to 1 or 0 to 1. The formula to scale input x to range [newx_max, newx_min] is given below where s(i) is normalized vector of specified range.

$$s(i) = (x(i) - \text{x\_min}) / (\text{x\_max} - \text{x\_min}) * (\text{newx\_max} - \text{newx\_min}) + \text{newx\_min}$$

b) Mean and Standard Deviation

Data is scaled so that it has zero mean and unity standard deviation. The formula for normalizing input x using standard deviation is given below.

$$s(i) = (x(i) - \text{mean\_x}) / \text{x\_standard\_deviation}$$

3.4.7.2 Sliding Window Approach

Time series data cannot be directly presented to neural network for training. To train neural network, training data should be in the form of (input, target) pair. To accomplish this task sliding window approach is adopted. In sliding a window, size and horizon is defined. Window size shows the number of input elements in each training pattern while horizon is the number of target elements in corresponding training pattern. Specified window slides over the time series, picking up (input, target) pairs.

Table 3.1 shows example dataset patterns after employing sliding window of size 12 having horizon one over the sample time series of 1960-1975. For simplicity we used years instead of years' data value. Values from row 1 to 12 represent input elements while value in row 13 represents the target value. The patterns obtained after applying sliding window would be equal to the size of time series less window size.

Number of Patterns = Size of Time series – Window Size

Table 3.1: Sample Time Series data 1960-1975 after employing window size of 12

	S. No.	Pattern1	Pattern2	Pattern3	Pattern4
Input units	1	1960	1961	1962	1963
	2	1961	1962	1963	1964
	3	1962	1963	1964	1965
	4	1963	1964	1965	1966
	5	1964	1965	1966	1967
	6	1965	1966	1967	1968
	7	1966	1967	1968	1969
	8	1967	1968	1969	1970
	9	1968	1969	1970	1971
	10	1969	1970	1971	1972
	11	1970	1971	1972	1973
	12	1971	1972	1973	1974
Target	13	1972	1973	1974	1975

3.4.7.3 Neural Network Generalization

The most common problem with backpropagation is overtraining or overfitting. To avoid overfitting early stopping method is used.

In this method data is divided in two subsets, training and validation. During training of neural network using training dataset, error on validation dataset is observed. The error on validation dataset usually decreases at start of training. Overfitting of model starts when error starts increasing on validation dataset. Early stopping stops the training when error on the validation set increases for specified number of iterations, and weights are adjusted at the iteration having minimum error on validation dataset [42]. Figure 3.9 shows error for training and validation dataset identifying stopping point.

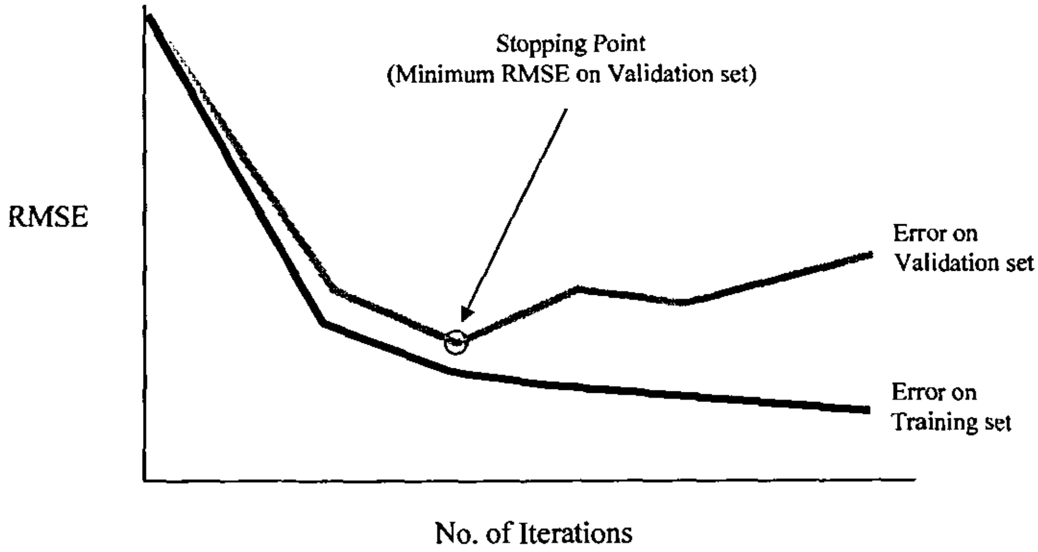

Figure 3.9: RMSE for validation and training set identifying Stopping point for training.

3.4.7.4 Number of Hidden Layer Neurons

Selecting correct number of hidden layer neurons is not a trivial task. There is no clear rule for determining the best number of hidden layer neurons, only heuristics are available. For example

a) Hidden layer neurons should not be more than twice the input layer neurons [43].

b) Hidden layer size should be in the range of input layer neurons and output layer neurons [44].

3.5 Model Fitness

To evaluate model fitness certain measures are required. For example analysis of actual and model predicted results using statistical functions i.e. Standard Deviation, Correlation Coefficient and Root Mean Square Error. These statistical terms are described below.

3.5.1 Standard Deviation

Standard deviation is a measure of variability. Standard deviation is defined as positive square root of variance [45]. Variance is sum of squared deviations about the sample mean Y_{mean} divided by n-1 [45]. The variance and standard deviation function for n samples of variable Y is given below.

$$Variance = s^2 = \frac{\sum_{i=1}^{n}(Y_i - Y_{mean})^2}{n-1}$$

$$S.D = s = \sqrt{\frac{\sum_{i=1}^{n}(Y_i - Y_{mean})^2}{n-1}}$$

3.5.2 Correlation Coefficient

Correlation is strength of relationship between variables. The measure of closeness between two variables is called coefficient of correlation [45]. If there are n independent pairs of values (X_i, Y_i), the sample correlation coefficient is defined as below.

$$Correlation\_Coefficient = r = \frac{\sum(X - X_{mean})(Y - Y_{mean})}{\sqrt{\sum(X - X_{mean})^2 \sum(Y - Y_{mean})^2}}$$

3.5.3 Root Mean Square Error (RMSE)

The root mean square error (RMSE) is defined as the square root of the mean squared difference between the forecast and actual values of the time series [46]. The RMSE function for n samples of variable Y is given below.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y\_predicted_i - Y\_actual_i)^2}$$

Chapter 4

Experimental Setup & Results

4. Experimental Setup & Results

This chapter presents in detail the experiments performed for prediction of monsoon rainfall using two Neural Network techniques, Backpropagation and LVQ. Different models were developed for this purpose by varying window size, number of hidden layer neurons and number of competitive neurons, in order to determine a suitable model.

This study is focused on Islamabad, capital city of Pakistan, that is one of the main gateway of monsoon in Pakistan and has higher variation among monsoon dominated areas of Pakistan over the period of 49 years from 1960-2008.

This chapter is organized as follows. In Section 4.1 experimental setup is presented including description of dataset, data preprocessing, early stopping criteria, evaluation criteria and initial network configuration. Section 4.2 presents the experimental results and gives a detailed analysis by comparing the monsoon prediction results of proposed neural network techniques with observed data and already in use statistical models. This Section gives details of selection of appropriate window size and hidden layer neurons. Prediction is made using the most suitable network configuration and results are evaluated for Monsoon prediction (2004-2008) for Islamabad and All Pakistan. Besides this experimental work, we also predicted the Summer Monsoon rainfall for Islamabad and All Pakistan for year 2009, which is given in Section 4.3.

4.1 Experimental Setup

4.1.1 Description of Dataset

Pakistan Meteorological Department has more than one hundred observatories across the country. This includes Automatic Weather Stations (AWS) and manual observing units for observing weather parameters e.g. rainfall, temperature, dew point temperature, wind speed and wind direction etc. These observations are taken on 03 hour basis, which are recorded in a register called pocket register and at the same time transmitted to regional meteorological offices in 15 to 30 minutes in decoded form for aviation use. CDPC (Computerized Data Processing Centre) Karachi is responsible for keeping and maintaining record of historical Meteorological Data. CDPC collects pocket registers from observatories on monthly basis and feeds data in ASCII file format into meteorological database.

The dataset obtained from Pakistan Meteorological Department for this study consists of 49 years (1960-2008) observed monthly rainfall data in millimeter (mm). The selected station data was available from 1960 onward.

Summer Monsoon (July, August and September) total rainfall was extracted from monthly rainfall. Islamabad Summer Monsoon (JAS) Rainfall is shown in Table 4.1 and Figure 4.1.

Table 4.1: Islamabad Summer Monsoon (JAS) Total Rainfall in mm (1960-2008)

Year	Rainfall	Year	Rainfall	Year	Rainfall	Year	Rainfall
1960	516.6	1973	867.0	1986	331.4	1999	711.3
1961	680.4	1974	744.8	1987	329.2	2000	717.0
1962	700.8	1975	780.6	1988	858.6	2001	898.0
1963	649.3	1976	1011.9	1989	763.1	2002	616.0
1964	402.6	1977	910.8	1990	931.7	2003	744.0
1965	243.6	1978	927.5	1991	727.5	2004	457.5
1966	439.4	1979	447.1	1992	819.6	2005	478.0
1967	660.9	1980	585.7	1993	457.4	2006	1111.6
1968	671.5	1981	1049.8	1994	1290.6	2007	1153.0
1969	409.2	1982	842.2	1995	1114.9	2008	726.0
1970	686.1	1983	1033.5	1996	729.9		
1971	635.2	1984	682.9	1997	849.2		
1972	235.9	1985	740.3	1998	872.4		

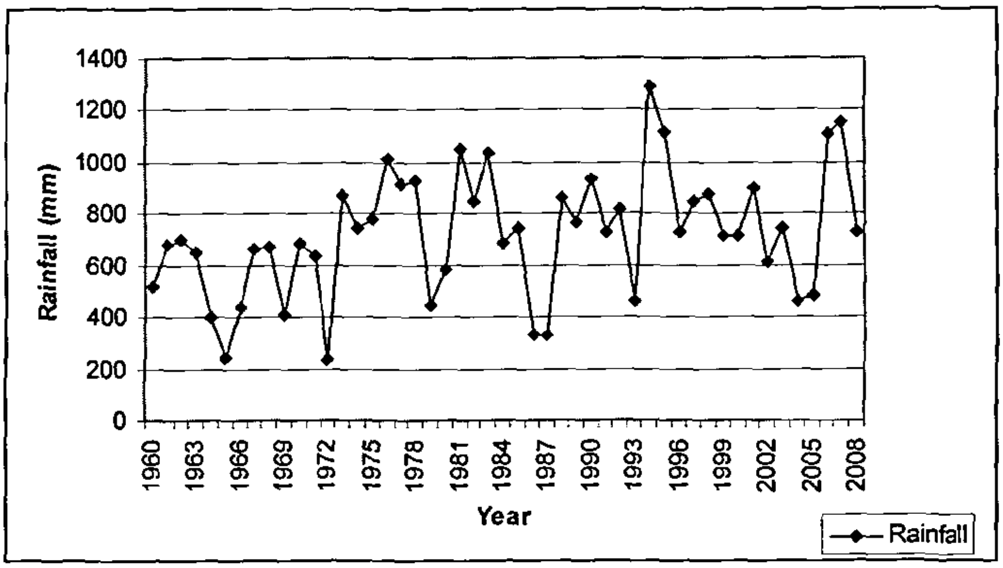

Figure 4.1: Islamabad Summer Monsoon (JAS) Total Rainfall (1960-2008)

4.1.2 Training and Test Datasets

Data is divided into two datasets training and testing. Forty three years data (1960-2003) is used as training dataset while five years data (2004-2008) is used as test dataset. Training dataset for each monsoon prediction from year 2004 to 2008 is given in Table 4.2. To utilize the data, training years were increased as prediction proceeded from 2004 to 2008. For example at the time of prediction for year 2005, year 2004 monsoon data was available; likewise at the time of prediction for year 2008, year 2007 monsoon data was available. We incorporated this data into training dataset to train the model on more recent patterns of rainfall.

Table 4.2 Training Dataset for prediction of Summer Monsoon Rainfall (2004-2008)

S. No.	Training Dataset	Predicting Year
1	1960-2003	2004
2	1960-2004	2005
3	1960-2005	2006
4	1960-2006	2007
5	1960-2007	2008

4.1.3 Early Stopping Criteria

As described in Chapter 3, the most common problem with backpropagation is over fitting or overtraining. To overcome this problem, early stopping method is used. For early stopping, data is divided into training and validation datasets. For our experiments validation patterns were chosen so that there is a validation set representative for each target class. So whenever more than two patterns belonged to same class, one pattern was selected as a validation pattern.

4.1.4 Evaluation Criteria

To determine model fitness, results were evaluated by comparing predicted monsoon rainfall with actual monsoon rainfall and calculating RMSE over 5 year's prediction. Correlation Coefficient was calculated between predicted and observed monsoon rainfall to see how well the proposed model captures the trend.

The predictive model is considered satisfactory if RMSE for prediction is less than the standard deviation of time series [15], and correlation is said to be excellent if correlation coefficient is great than 0.80.

4.1.5 Data Preprocessing

i) Data Discretization

Data is classified in bands of 50mm for improving the efficiency of neural network learning, as continuous data may add noise and degrade the performance of neural network learning. We used band of 50mm as it is acceptable range when predicting monsoon total rainfall. For example it would be appropriate and more practical to say that coming year monsoon total rainfall would be in range from 400mm to 450 mm, rather than talking about exact value 412mm or 435mm, which is insignificant on seasonal level.

Table 4.3 shows the class labels for each rainfall band after discretization. Rainfall range was selected from 0.0mm to 1300.0mm as monsoon rainfall at most of the stations in Pakistan persists in this range.

Table 4.3: Class labeling for each band of rainfall

Rainfall range (mm)			Class	Rainfall range (mm)			Class
0	-	50	1	650	-	700	14
50	-	100	2	700	-	750	15
100	-	150	3	750	-	800	16
150	-	200	4	800	-	850	17
200	-	250	5	850	-	900	18
250	-	300	6	900	-	950	19
300	-	350	7	950	-	1000	20
350	-	400	8	1000	-	1050	21
400	-	450	9	1050	-	1100	22
450	-	500	10	1100	-	1150	23
500	-	550	11	1150	-	1200	24
550	-	600	12	1200	-	1250	25
600	-	650	13	1250	-	1300	26

ii) Data Normalization

Data is normalized before presenting it to the network for training. To normalize the data we applied, mean & standard deviation normalization method.

4.1.6 Backpropagation Network Configuration

Neural network configuration is as below.

Learning rate α	=	0.10
Hidden Layer Activation function	=	Logistic Sigmoid
Output Layer Activation function	=	Logistic Sigmoid

Activation function used for hidden and output layer is logistic sigmoid, which is the commonly used activation function [37]. It is also called binary sigmoid function because it ranges output from 0 to 1 [37].

4.2 Results and Analysis

4.2.1 Experiments using Backpropagation

Selection of suitable window size and appropriate number of hidden layer neurons is a challenging task. These parameters are best determined empirically and there is no clear and specific criterion for selection of these parameters [32]. There are some heuristics given in Chapter 3, but these heuristics are just a starting point. To obtain a suitable window size and appropriate number of hidden layer neurons, for summer monsoon rainfall prediction, several experiments were conducted to train the model using window sizes from 04 to 12 and number of hidden layer neurons from 03 to 35 for each window size. Window size selection from 04 to 12 is in compliance with work of Guhathakurta [15], who selected window sizes from 05 to 12 and achieved suitable results on window sizes 11 and 12 for Indian Monsoon Rainfall Prediction.

RMSE was calculated for prediction of Islamabad Summer Monsoon rainfall (2004 – 2008) using window sizes from 04 to 12 and hidden layer neurons from 03 to 35. Table 4.4 and Figure 4.2 shows the minimum RMSE obtained at each window size with number of hidden layer neurons.

It is evident from Table 4.4 and Figure 4.2 that window size 12 is an appropriate window size which produced minimum RMSE 2.65 among all window sizes using 08 number of hidden layer neurons.

Table 4.4: Minimum RMSE (Class) of each window size (04 to 12) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall with number of hidden layer neurons (03 to 35).

Window Size	Minimum RMSE	No. of Hidden Layer Neurons
04	4.80	16
05	4.71	11
06	3.19	10
07	4.34	11
08	4.47	08
09	3.44	33
10	3.82	09
11	3.74	04
12	2.65	08

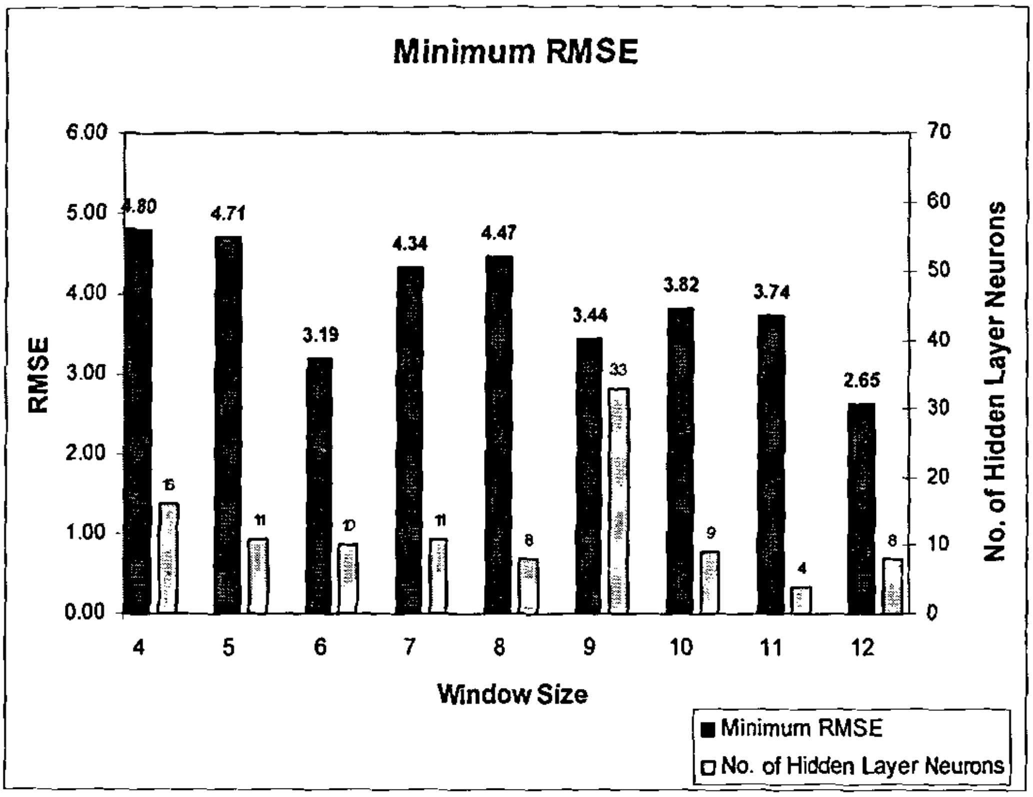

Figure 4.2: Minimum RMSE (Class) of each window size (04 to 12) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall with hidden layer neurons from 03 to 35 using backpropagation.

Figure 4.3 shows RMSE at window size 12 using hidden layer neurons from 03 to 35. Figures depicting RMSE obtained on window sizes from 04 to 11 are given in Appendix A.1. It is obvious from Figure 4.3 that hidden neurons 08 have minimum RMSE.

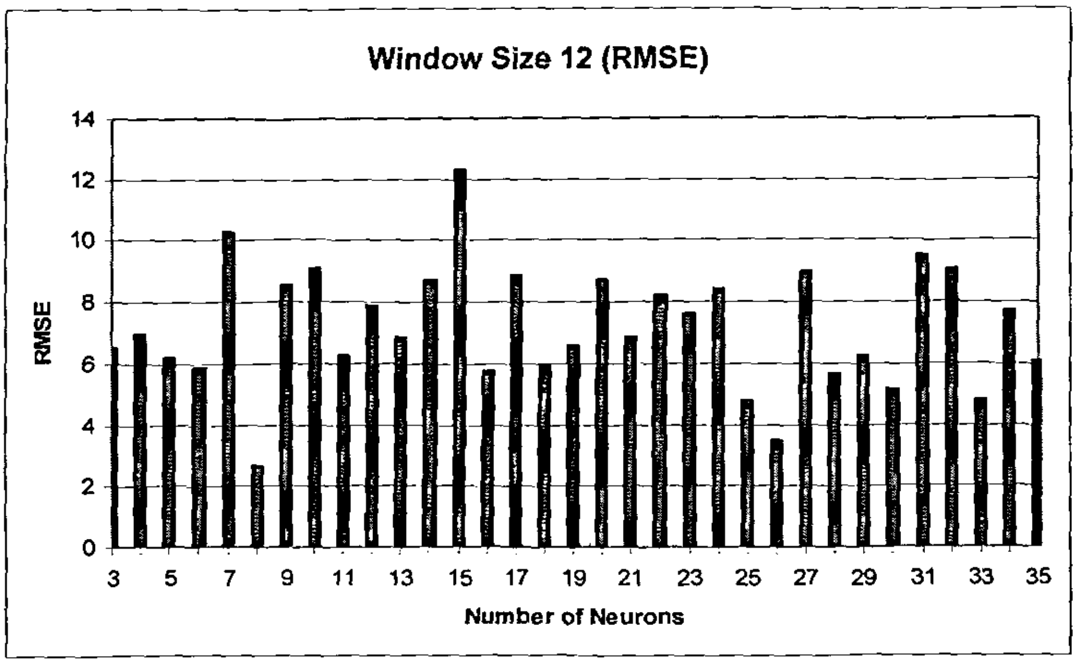

Figure 4.3: RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using window size 12 and number of hidden layer neurons from 3 to 35.

4.2.2 Islamabad Monsoon Rainfall Prediction from 2004 to 2008 using Backpropagation

Islamabad monsoon rainfall prediction (2004-2008) was made using backpropagation with window size 12 and hidden layer neurons 08. Actual and Predicted summer monsoon rainfall is given in Table 4.5 and Figure 4.4 below. RMSE obtained over the prediction of these 05 years was 2.65, which was almost half of the standard deviation for Islamabad Monsoon Rainfall time series and thus considered to be satisfactory as per evaluation criteria given in Section 4.1.5. Correlation coefficient achieved was 0.90 between actual and predicted rainfall. This shows the strength of the model.

Table 4.5: Islamabad Predicted Summer Monsoon Rainfall (mm) using window size 12 and Hidden Layer Neurons 8

	2004	2005	2006	2007	2008
Predicted	300-350	700-750	1100-1150	1100-1150	700-750
Actual	450-500	450-500	1100-1150	1150-1200	700-750

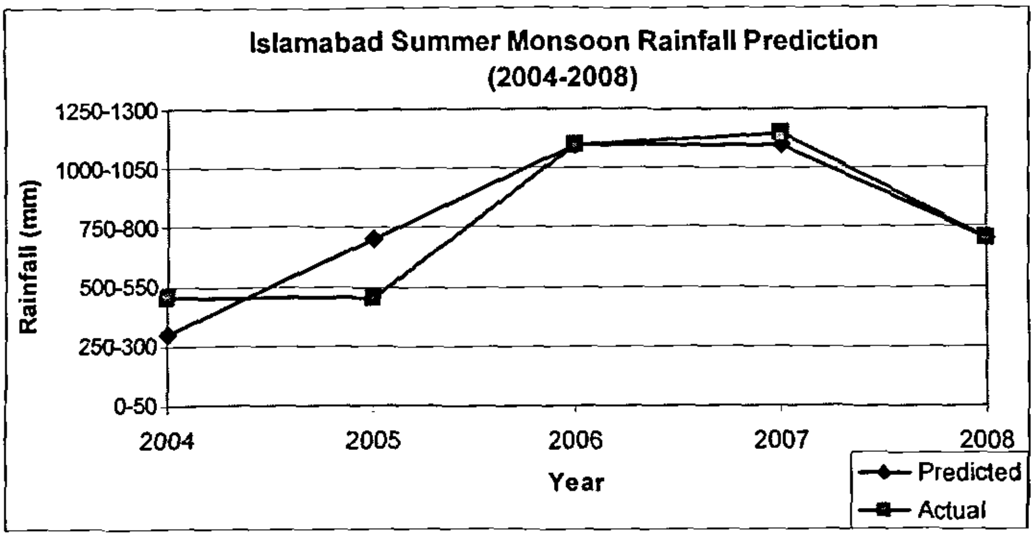

Figure 4.4: Islamabad Summer Monsoon Rainfall Prediction (2004-2008) using window size 12 and hidden layer neurons 08.

4.2.3 Experiments using LVQ

Several experiments were conducted using LVQ for obtaining suitable window size and appropriate number of competitive layer neurons. LVQ network was initialized using learning rate 0.01. We fixed the number of iterations to 300, as after initial experiments we found that increasing the number of iterations further does not lead to any performance improvement.

Through experiments, we found that window size 12 with 25 number of competitive neurons is suitable window size having minimum RMSE 2.72 among all window sizes from 04 to 12 and competitive layer neurons from 03 to 35. Table 4.6 and Figure 4.5 show the minimum RMSE obtained at each window size with number of competitive layer neurons.

Table 4.6: Minimum RMSE (Class) of each window size (4 to 12) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall with number of competitive layer neurons using LVQ.

Window Size	Minimum RMSE	No. of Hidden Layer Neurons
04	4.82	34
05	3.87	17
06	2.83	19
07	2.90	21
08	3.74	13
09	4.17	12
10	4.31	04
11	4.07	32
12	2.72	25

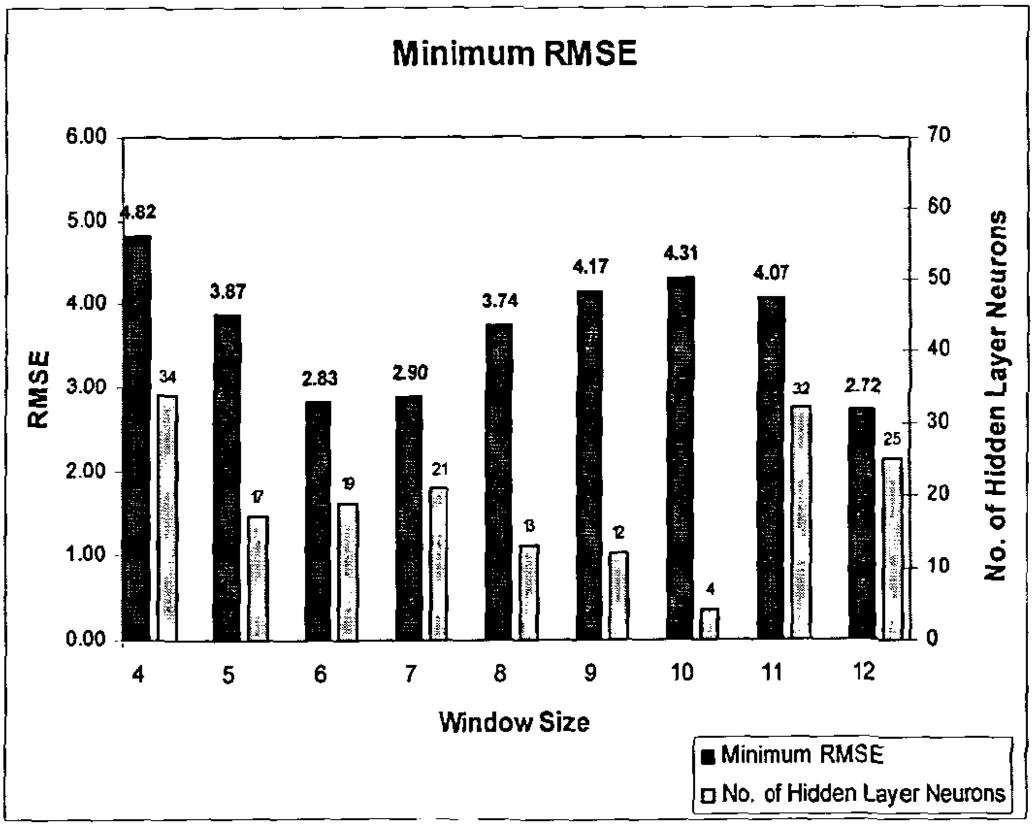

Figure 4.5: Minimum RMSE (Class) of each window size (4 to 12) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall with competitive layer neurons from 03 to 35 using LVQ.

RMSE for window size 12 using competitive neurons 03 to 35 is given in Figure 4.6. RMSE for window sizes from 04 to 11 are given in Appendix A.2. It is obvious from Figure 4.6 that 25 hidden layer neurons have minimum RMSE.

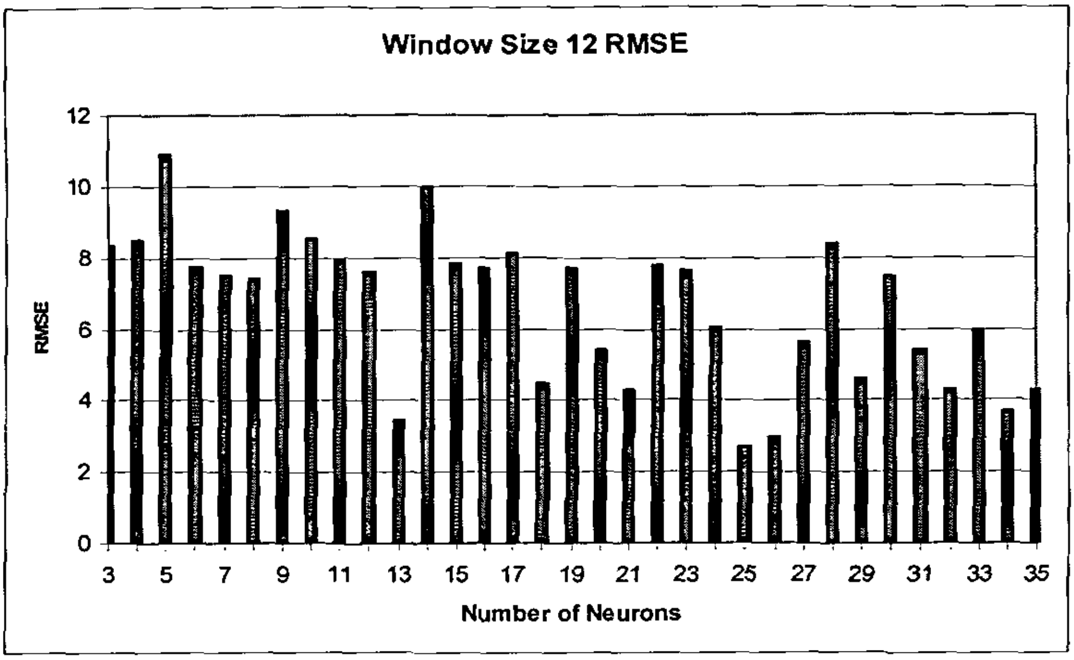

Figure 4.6: RMSE (Class) on prediction of 12 years (2004-2008) Islamabad Summer Monsoon Rainfall using window size 4 and number of hidden layer neurons from 3 to 35.

4.2.4 Islamabad Monsoon Rainfall Prediction from 2004 to 2008 using LVQ

Islamabad Monsoon Rainfall prediction was made using window size 12 and competitive neurons 25. We compared the predicted results obtained using LVQ with actual monsoon rainfall and found strong correlation coefficient of 0.95 between actual and predicted values having RMSE 2.72 that is in acceptable range being almost half of the standard deviation as described in Section 4.1.5. Actual and predicted values of Islamabad Monsoon rainfall are illustrated in Table 4.7 and Figure 4.7 below.

Table 4.7: Islamabad Predicted Summer Monsoon Rainfall (mm) using window size 12 and Hidden Layer Neurons 25

	2004	2005	2006	2007	2008
Predicted	300-350	300-350	1250-1300	1000-1050	750-800
Actual	450-500	450-500	1100-1150	1150-1200	700-750

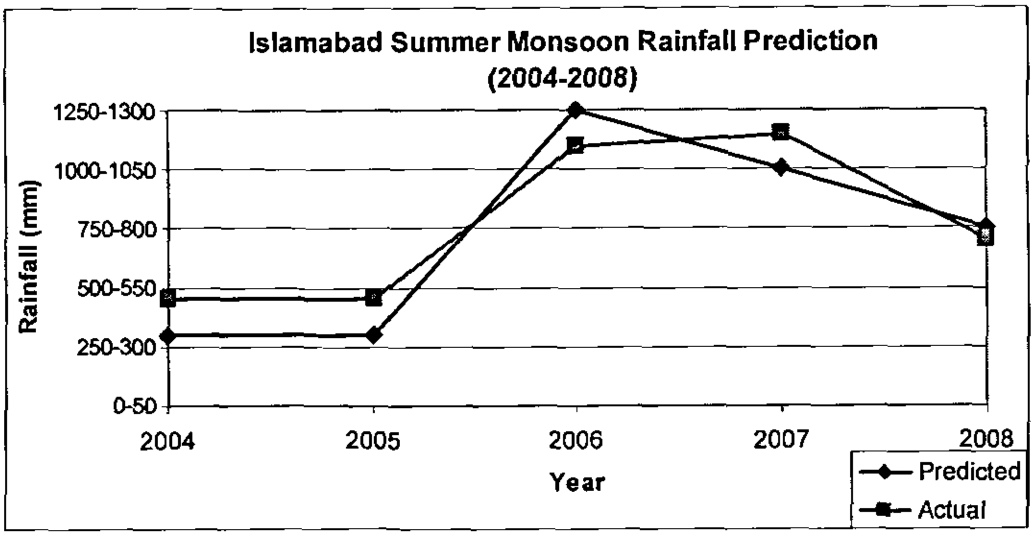

Figure 4.7: Islamabad Summer Monsoon Rainfall Prediction (2004-2008) using window size 12 and competitive neurons 25.

4.2.5 Comparison: Backpropagation vs LVQ for Islamabad Monsoon Rainfall Prediction

We employed both Backpropagation and LVQ techniques for monsoon rainfall prediction and found that both techniques are capable of predicting monsoon rainfall having above 90 percent correlation, and RMSE with almost half of the standard deviation. A comparison of predicted monsoon rainfall using both techniques with actual rainfall is given in Table 4.8 and Figure 4.8.

Table 4.8: Islamabad Actual and Predicted Rainfall using Backpropagation and LVQ

	2004	2005	2006	2007	2008
LVQ	300-350	300-350	1250-1300	1000-1050	750-800
Backpropagation	300-350	700-750	1100-1150	1100-1150	700-750
Actual	450-500	450-500	1100-1150	1150-1200	700-750

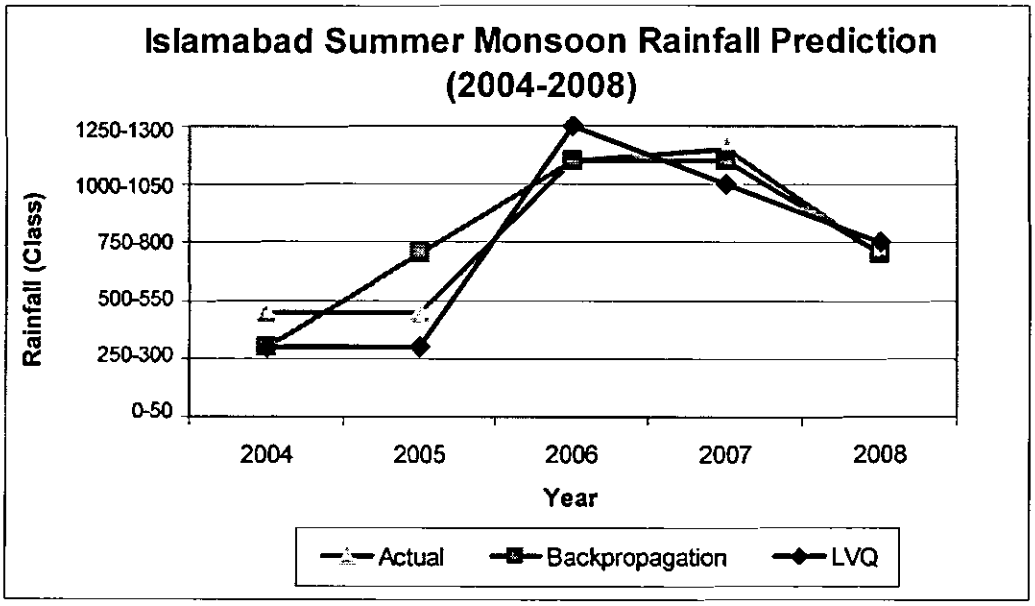

Figure 4.8: Islamabad Monsoon Actual and predicted Rainfall (mm) using LVQ & Backpropagation

RMSE

A comparison of RMSE and correlation coefficient calculated for Islamabad Monsoon Rainfall prediction (2004-2008) using BP and LVQ is given in Table 4.9.

Table 4.9: RMSE and Correlation Coefficient using Backpropagation and LVQ for Islamabad Monsoon Rainfall prediction (2007-2008)

Technique	RMSE	Correlation Coefficient
Backpropagation	2.65	0.90
LVQ	2.72	0.95
Standard Deviation is 4.84		

It is evident from Table 4.9 that RMSE for both techniques was less than the standard deviation. This indicates model fitness.

Although our experimental results show that both techniques give satisfactory results, it should be noted that Backpropagation networks are general in nature and can be used in almost every field that uses neural networks for problem solving [37]. On the other hand LVQ networks are supervised competitive neural networks that are designed for classification problems where only one out of several outputs is on [37].

LVQ overcomes the problem that we might face in Backpropagation of having output 1 for more than one output neurons. This problem may raise confusion in deciding the correct class, although we have not faced this problem in our study but it may raise potential problems.

LVQ takes less training time than Backpropagation [47]. In our case of monsoon prediction almost a year in advance, training time difference that was in seconds is insignificant.

Thus we may conclude that both techniques are capable of producing satisfactory results and can be adopted as operational models for monsoon rainfall prediction.

4.2.6 Comparison: Backpropagation vs LVQ vs Statistical Downscaling

Pakistan Meteorological Department started monsoon prediction using statistical downscaling technique in 2007. We performed a comparison of Islamabad Monsoon rainfall for year 2007 and 2008 because prediction from statistical downscaling model was only available for year 2007 and 2008.

For comparison actual rainfall was used as reference value. Predicted rainfall from both neural network techniques and statistical downscaling technique was compared with its corresponding actual monsoon rainfall.

We discretized the predicted values of statistical downscaling technique and actual rainfall in 50mm band as well for comparison. Table 4.10 shows RMSE (mm in terms of class) and Correlation Coefficient for prediction of year 2007 and 2008 using ANN, LVQ and statistical Downscaling technique.

Table 4.10: RMSE and Correlation Coefficient using ANN, LVQ and Statistical Downscaling for Islamabad Monsoon Rainfall prediction (2007-2008)

Technique	RMSE	Correlation Coefficient
Backpropagation	0.7	1.0
LVQ	2.2	1.0
Statistical Downscaling	5.8	-1.0
Standard Deviation is 4.84		

It is evident from Table 4.10 that neural network techniques (Backpropagation and LVQ) performed much better than statistical downscaling technique as backpropagation gave only 0.7 RMSE with correlation coefficient 1.0 while statistical downscaling technique produced much higher RMSE 5.8 with correlation coefficient -1.0. Though LVQ RMSE 2.2 was higher than backpropagation but even then it remained much less than Statistical Downscaling technique and less than half of the standard deviation.

Thus we may conclude that both neural network techniques outperform the statistical downscaling model for monsoon rainfall prediction.

4.2.7 All Pakistan Summer Monsoon Rainfall Prediction

Pakistan is divided to 56 Meteorological regions. Pakistan Meteorological Department issues All Pakistan Area Weighted Monsoon rainfall prediction using statistical linear regression model.

We employed both neural network techniques backpropagation and LVQ for PSMR prediction from 2004 to 2008. Here we used training dataset from 1961 to 2003 because the dataset at All Pakistan level was available from 1961 onward.

We trained the model finding suitable window size and hidden layer neurons by attempting window sizes from 04 to 12 and hidden layer neurons from 03 to 35. Results obtained using both neural network techniques are given below.

4.2.8 PSMR prediction using Backpropagation

PSMR prediction is made using same neural network configuration as given in Section 4.1.7 for the case of Islamabad. Suitable window size and number of hidden layer neurons was obtained empirically. Table 4.11 and Figure 4.9 below shows the minimum RMSE at each window size with number of hidden layer neurons. Window sizes 08 and 06 have minimum RMSE 0.45 with hidden layer neurons 16 and 28 respectively. As window size 08 gave the same minimum RMSE 0.45 with fewer number of hidden layer neurons 16, we will consider it suitable window size. Figures containing details for each window size with corresponding number of hidden layer neurons are given in Appendix A.3.

Table 4.11: Minimum RMSE (Class) of each window size (04 to 12) for prediction of 5 years (2004-2008) Pakistan Summer Monsoon Rainfall with number of hidden layer neurons (03 to 35) using Backpropagation.

Window Size	Minimum RMSE	No. of Hidden Layer Neurons
04	0.63	5
05	0.63	10
06	0.45	28
07	0.77	19
08	0.45	16
09	0.77	24
10	0.63	21
11	0.63	30
12	0.77	31

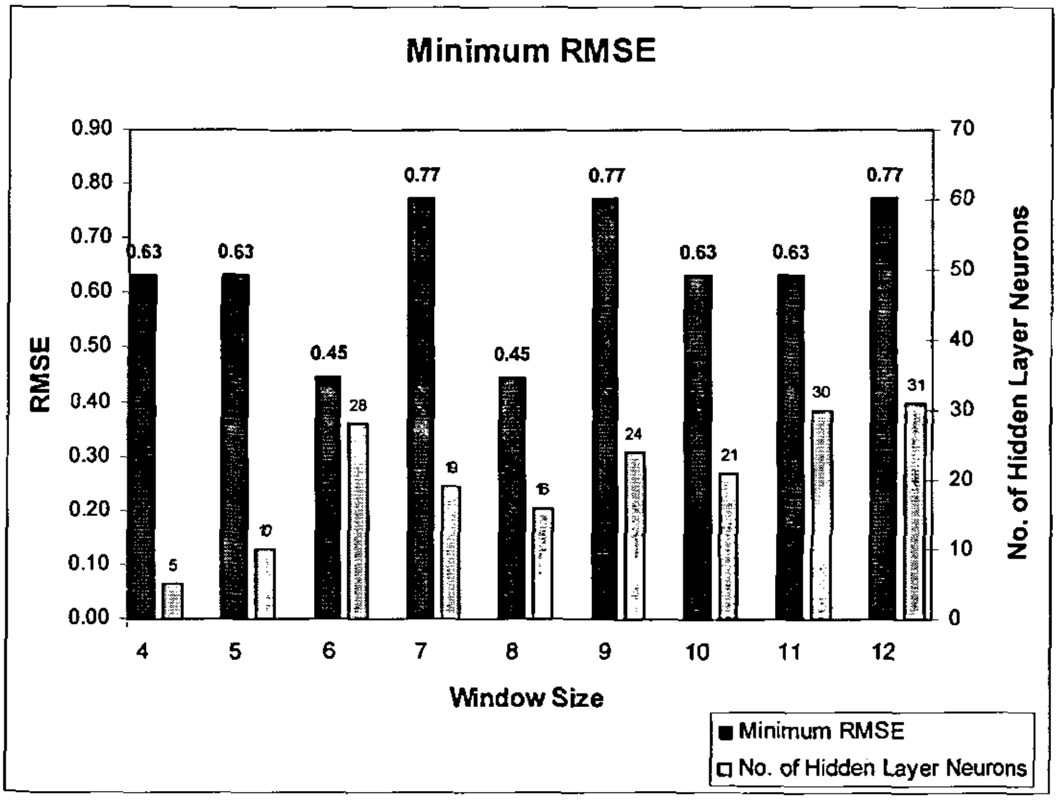

Figure 4.9: Minimum RMSE (Class) of each window size (4 to 12) for prediction of 5 years (2004-2008) Pakistan Summer Monsoon Rainfall with hidden layer neurons from 03 to 35 using Backpropagation.

Actual and Predicted Pakistan Monsoon Rainfall using window size 8 and hidden layer neurons 16 is given in Table 4.12 and Figure 4.10 below.

Table 4.12: Pakistan Predicted and Actual Summer Monsoon Rainfall (in terms of class) 2004-2008 using optimum window size 8 and Hidden Layer Neurons 16

	2004	2005	2006	2007	2008
Predicted	51-100	101-150	201-250	101-150	101-150
Actual	51-100	101-150	151-200	101-150	101-150

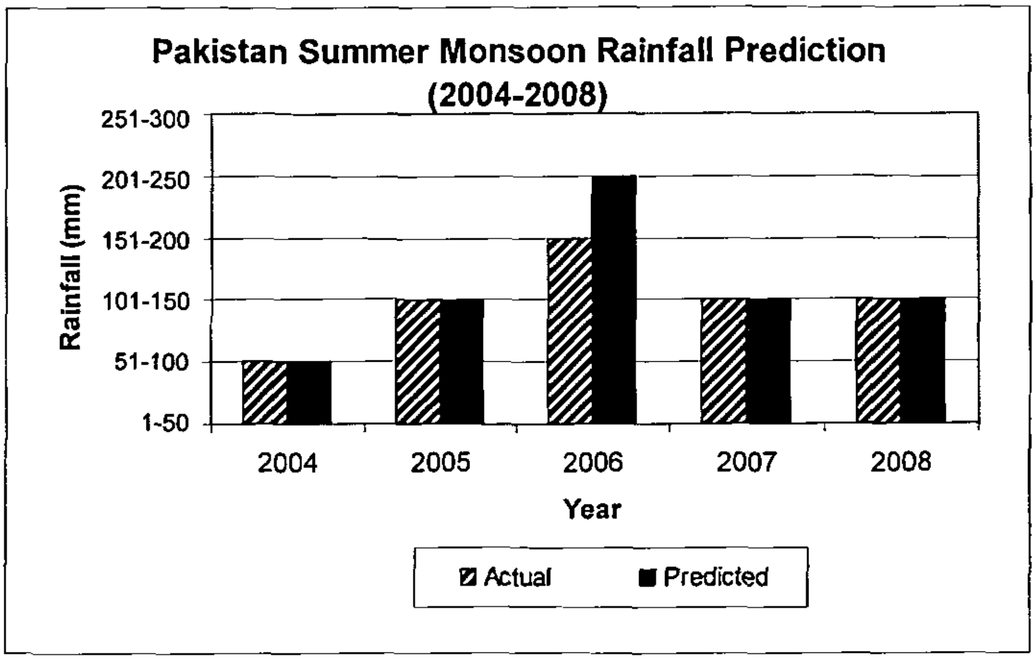

Figure 4.10: Minimum RMSE (Class) of each window size (4 to 12) on prediction of 5 years (2004-2008) Pakistan Summer Monsoon Rainfall with hidden layer neurons from 03 to 35 using Backpropagation.

4.2.9 PSMR prediction using LVQ

LVQ was employed using window sizes 12 to 04 and competitive layer neurons from 03 to 35 to find suitable window and competitive layer size for PSMR prediction. Table 4.13 and Figure 4.11 below illustrate the minimum RMSE of each window size with hidden layer neurons. It is obvious from Table 4.13 and Figure 4.11 below, that window size 04 gave smaller RMSE 0.77 among all window sizes using hidden layer neurons 05. Figures containing details for each window size with corresponding number of competitive layer neurons are given in Appendix A.4.

Table 4.13: Minimum RMSE obtained at each window size with Number of Competitive Layer Neurons

Window Size	Minimum RMSE	No. of Hidden Layer Neurons
04	0.77	05
05	1.10	06
06	1.10	07
07	1.10	06
08	1.18	05
09	1.18	06
10	1.41	05
11	1.48	15
12	1.73	11

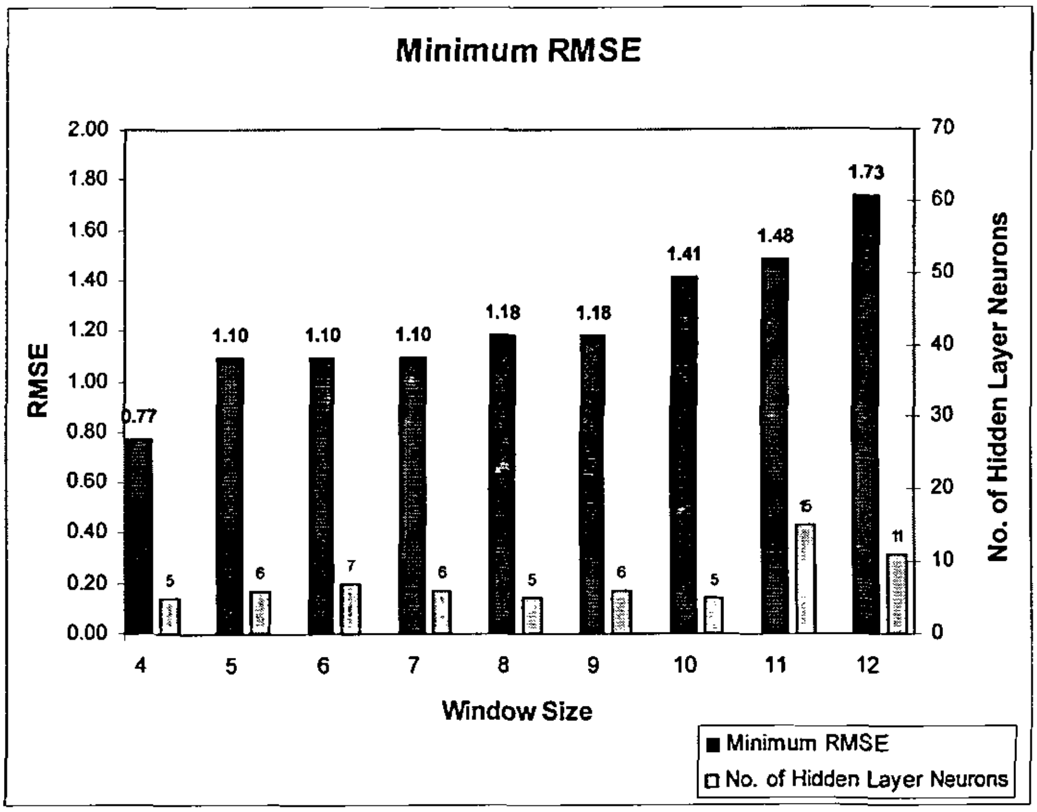

Figure 4.11: Minimum RMSE (Class) of each window size (4 to 12) for prediction of 5 years (2004-2008) Pakistan Summer Monsoon Rainfall with competitive layer neurons from 03 to 35 using Backpropagation.

4.2.10 Comparison: Backpropagation vs LVQ vs Statistical model for PSMR prediction

For evaluation, we not only compared the predicted results with actual rainfall but with the predicted results of linear regression model that is the operational monsoon rainfall prediction model of Pakistan Meteorological Department.

Table 4.14 and Figure 4.12 below present the predicted values of Neural Network Models Backpropagation, LVQ and PMD's statistical model (linear regression).

Table 4.14: Pakistan Actual and Predicted Rainfall using Backpropagation, LVQ and Statistical model.

	2004	2005	2006	2007	2008
LVQ	51-100	51-100	101-150	51-100	101-150
BP	51-100	101-150	201-250	101-150	101-150
Statistical	151-200	201-250	201-250	151-200	201-250
Actual	51-100	101-150	151-200	101-150	101-150

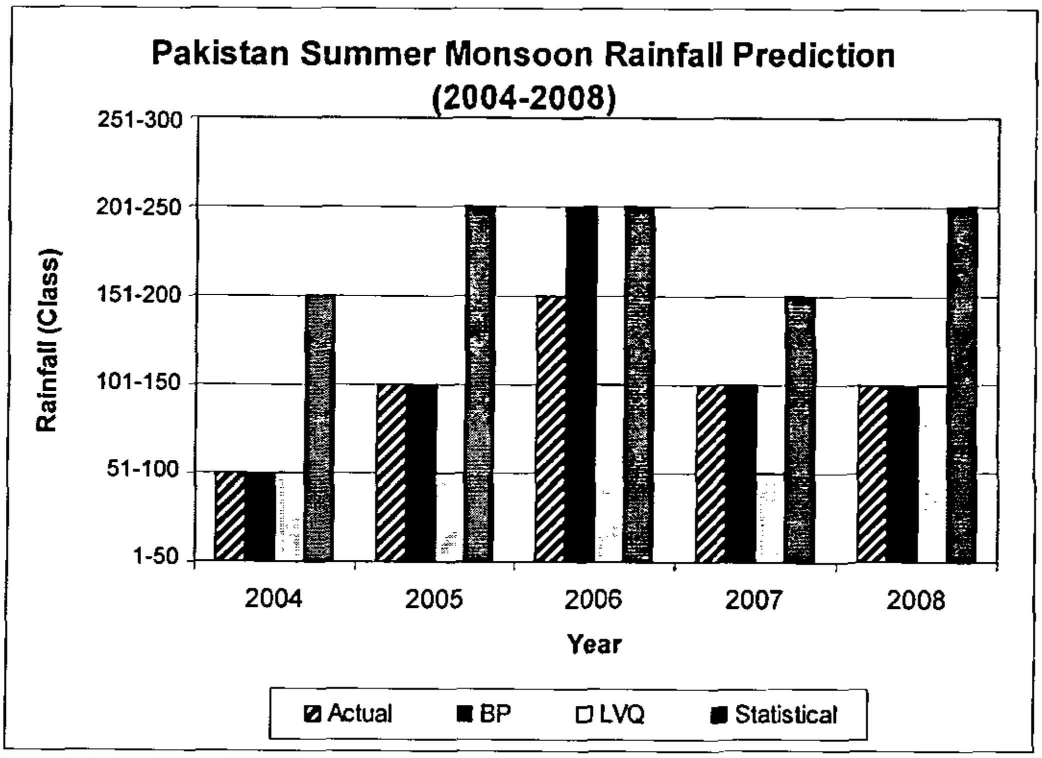

Figure 4.12: Pakistan Monsoon Rainfall Actual and predicted using LVQ, BP & Statistical Model.

RMSE

RMSE produced by different techniques for predicting All Pakistan Summer Monsoon Rainfall is given in Table 4.15 with corresponding Correlation Coefficient.

Backpropagation gave RMSE 0.45 that is less than half of the standard deviation 1.08, while statistical model gave RMSE of 1.70. Correlation coefficient using backpropagation is 0.97, which is much better than that of statistical technique which is 0.65. LVQ produced RMSE 0.77 that is greater than error produced by Backpropagation, however it is much smaller than the error of statistical model that is 1.70 and falls below the standard deviation. These results show the strength of neural network in Pakistan Monsoon Rainfall prediction over traditional statistical technique.

Table 4.15: RMSE and Correlation Coefficient using ANN, LVQ and Statistical Technique for All Pakistan Monsoon Rainfall prediction (2004-2008)

Technique	RMSE	Correlation Coefficient
BP	0.45	0.97
LVQ	0.77	0.65
Statistical	1.70	0.65
Standard Deviation is 1.08		

4.3 All Pakistan and Islamabad Monsoon Rainfall Prediction for coming Year 2009 using Neural Network

On the basis of optimum neural network models developed during this study, we made a prediction for coming year monsoon rainfall. The prediction for Islamabad and All Pakistan Summer Monsoon Rainfall using Backpropagation and LVQ is given below in Table 4.16.

Table 4.16: Summer Monsoon (JAS) Rainfall Prediction for year 2009

Area	Monsoon Rainfall Prediction By	
	Backpropagation	LVQ
All Pakistan (Area Weighted)	200mm – 250mm	150mm – 200mm
Islamabad	700mm – 750mm	700mm – 750mm

This prediction is being made in month of May 2009, as we carried out our study this month. The data for training and prediction of proposed models becomes available on 1st of October. Proposed models can predict coming year monsoon rainfall with lead time of 09 months. It is longer lead time as compared to PMD's operational statistical model that predicts coming monsoon rainfall with only 01 month lead time.

4.4 Summary

In this chapter we investigated two neural network techniques LVQ and Backpropagation with different configurations of window size and hidden/competitive layer neurons for prediction of Islamabad and All Pakistan Summer Monsoon Rainfall (2004-2008). Window size 12 found to be suitable window size for both models for Islamabad monsoon rainfall prediction. In the case of PSMR prediction, satisfactory results were obtained on windows sizes 08 and 04 using backpropagation and LVQ respectively. Results of these models for monsoon rainfall prediction were evaluated by comparison with actual rainfall as well as with monsoon prediction of PMD's Statistical Model and Statistical Downscaling model. Backpropagation and LVQ both performed equally well having RMSE much less than standard deviation as compared to statistical downscaling and statistical models, which produced much higher RMSE.

Proposed models use only historical monsoon rainfall data and are not dependent on any other parameters like statistical models which requires number of parameters e.g. South American Pressure departure, Himalayas Snow accumulation in the month of May, Temperature at Lahore, Islamabad, Sialkot, Multan for the months of April and May. Statistical downscaling model also requires number of parameters' output e.g. temperature, sea level pressure and precipitation all over the globe from GCM and correlates them with historical monsoon rainfall to find predictands for summer monsoon rainfall prediction.

The dynamical models require number of hours on High Performance Clusters having computational power of Tera Flops to simulate the model. Our proposed model requires few seconds on desktop PC for training and prediction.

Thus the proposed neural network models are better than traditional statistical and statistical downscaling models not only in term of accuracy but also in terms of resources that are required for prediction.

Chapter 5

Conclusion & Future Work

5. Conclusion & Future Work

In this study we explored the use of Data Mining techniques i.e. neural networks, to develop models for PSMR prediction. Summary of our research contributions is as follows.

i) We developed our models by employing two well known neural network techniques i.e. Backpropagation and LVQ. Using these models, we predicted monsoon rainfall for Islamabad, the capital city of Pakistan. The models were also used for All Pakistan summer monsoon rainfall prediction. We compared the results of both techniques with actual monsoon rainfall data for prediction of 5 years (2004-2008) monsoon rainfall. RMSE of backpropagation was 2.65 and 0.45 for Islamabad and All Pakistan monsoon rainfall prediction respectively that is almost half of the standard deviation of monsoon rainfall over 49 years from 1960 to 2008. For LVQ the RMSE was 2.72 and 0.77 for Islamabad and All Pakistan respectively, that is less than the standard deviation. This reveals that both techniques have capability of predicting PSMR with similar accuracy.

ii) Configuration of neural network parameters to achieve satisfactory results is a challenging task. To obtain suitable model configuration we employed different variations of both techniques using window sizes from 04 to 12 and hidden/competitive layer neurons from 03 to 35. Suitable configuration of backpropagation for Islamabad monsoon rainfall prediction was achieved using window size 12 with 08 hidden layer neurons. For All Pakistan monsoon rainfall prediction, suitable configuration of backpropagation was obtained using window size 08 and hidden layer neurons 16. On the other hand LVQ results were appropriate for Islamabad monsoon rainfall prediction when used with window size 12 and competitive neurons 25. For All Pakistan, LVQ network gave appropriate results using window size 04 and competitive neurons 05.

iii) We used real monsoon rainfall data from 1960 to 2008 for training of the models. The approach of using only previous years' monsoon rainfall data allows proposed models to predict coming year monsoon rainfall with longer lead time of 09 months. On the other hand PMD's regression based statistical model and statistical downscaling model is dependent on number of variables that limits their ability of predicting monsoon rainfall beyond one month lead time.

iv) We performed comparison of proposed models' results with the results of PMD's regression based statistical model for All Pakistan monsoon rainfall prediction (2004-2008). RMSE obtained using PMD's statistical model is 1.7, which is not only much higher than the RMSE of both neural network techniques i.e. Backpropagation 0.45 and LVQ 0.77, but also higher than the standard deviation. We also performed comparison of proposed models' results with the results of statistical downscaling model for Islamabad Monsoon rainfall prediction. Comparison is made for year 2007 and 2008, because prediction of statistical downscaling technique was available from 2007 onward. The RMSE obtained using statistical downscaling is 5.8, which is much higher than that of both neural networks prediction errors i.e. RMSE 0.7 of BP and 2.2 of LVQ, and also higher than the standard deviation. Comparison reveals the strength of proposed models over regression based statistical model and statistical downscaling model for prediction of PSMR.

This study reveals that proposed neural network models have the capability of predicting PSMR accurately with longer lead time using fewer resources and hence can be deployed as operational monsoon rainfall prediction models.

5.1 Future Work

i) We focused this study for prediction of summer monsoon rainfall for Islamabad, in future this study may be extended to cover other monsoon dominated areas of Pakistan e.g. Jhelum, Sialkot and Lahore.

ii) The focus of this study was summer monsoon rainfall. This study can be replicated to predict Pakistan winter rainfall that has impact on winter crops.

iii) In this study we projected seasonal rainfall. Daily rainfall has also great importance for daily activities e.g. outdoor events, construction, where the major concern is whether it is going to rain tomorrow or in next few days or not. Proposed techniques using previous years' daily rainfall data may be applied to predict rainy days.

iv) Onset of monsoon in Pakistan i.e. when monsoon is likely to commence in Pakistan next year, has great importance. Work may be carried out to predict onset of coming monsoon using historical monsoon onset data.

v) In this study we explore the use of Backpropagation and LVQ for PSMR prediction. There is scope of investigating other data mining techniques for this problem i.e. Support Vector Machines, Genetic Algorithms and Decision Trees.

Appendix A

RMSE using Backpropagation & LVQ

Appendix A: RMSE using Backpropagation & LVQ

A.1 RMSE using Backpropagation for Islamabad Summer Monsoon Rainfall Prediction

Figures below from Figure A.1 to Figure A.12 depict RMSE for Islamabad Monsoon Rainfall Prediction (2004-2008) using Backpropagation with window size (04 to 12) and hidden layer neurons (03 to 35).

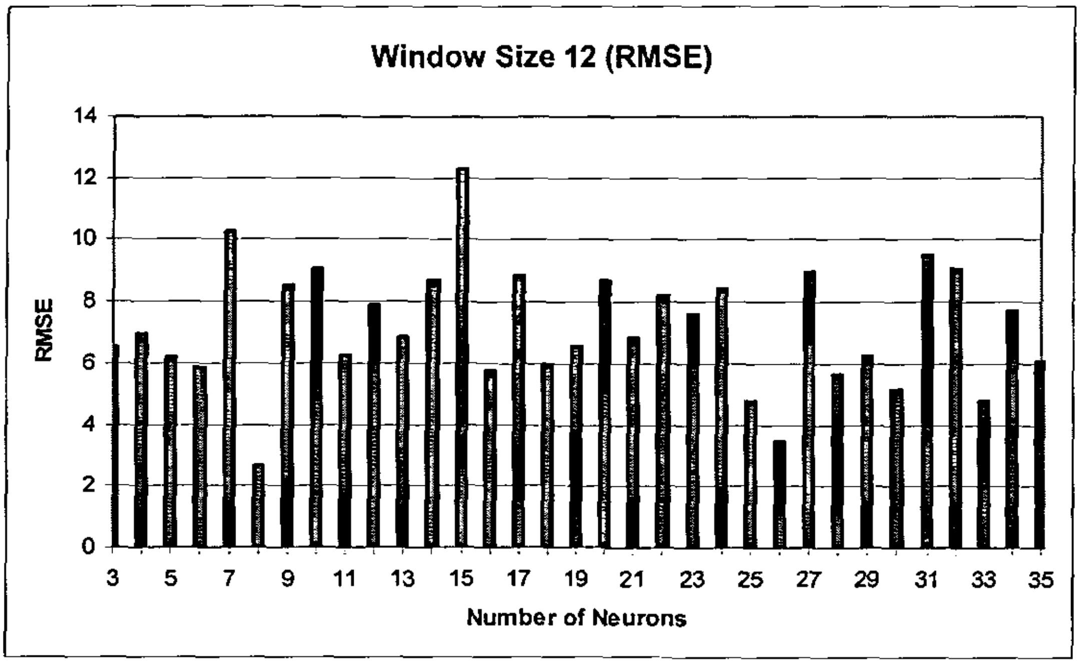

Figure A.1 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 12 and number of hidden neurons from 3 to 35.

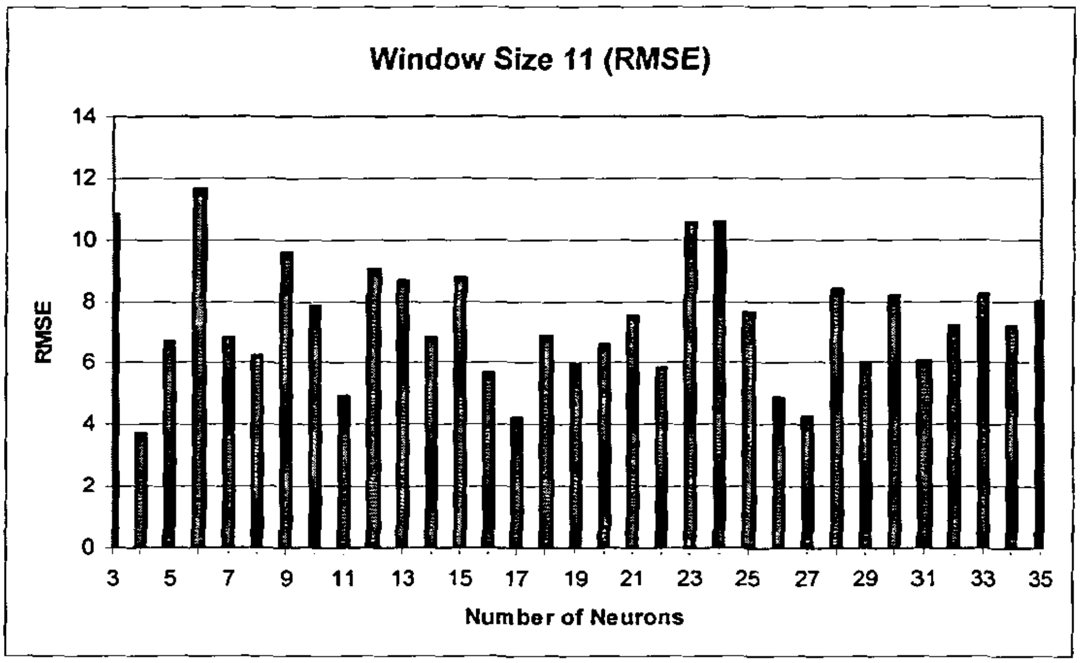

Figure A.2 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 11 and number of hidden neurons from 3 to 35.

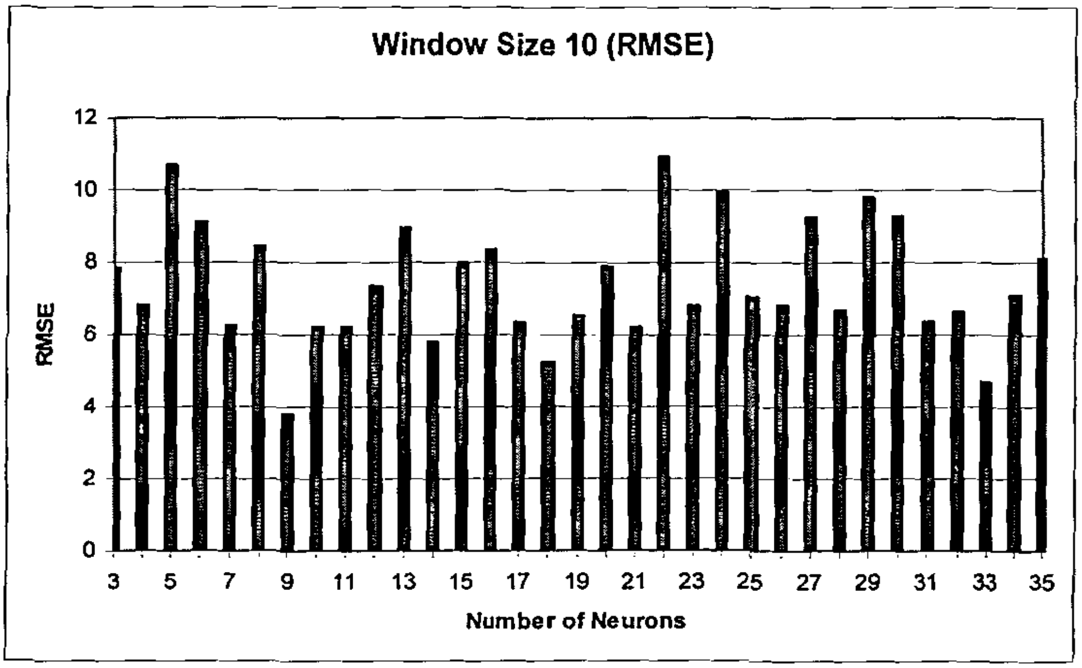

Figure A.3 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 10 and number of hidden neurons from 3 to 35.

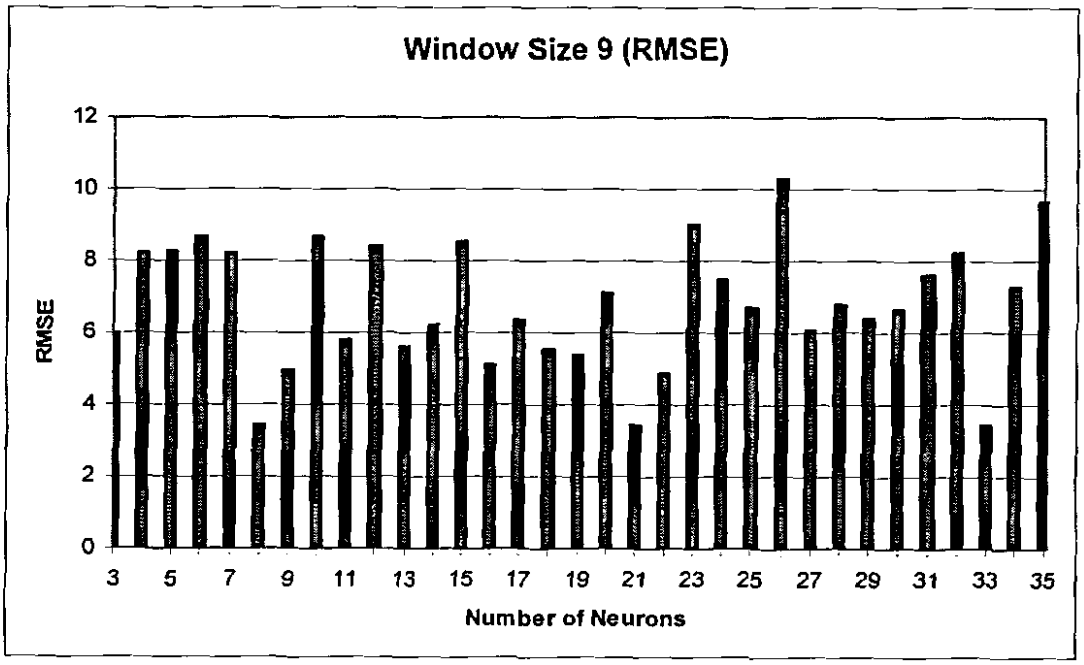

Figure A.4 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 9 and number of hidden neurons from 3 to 35.

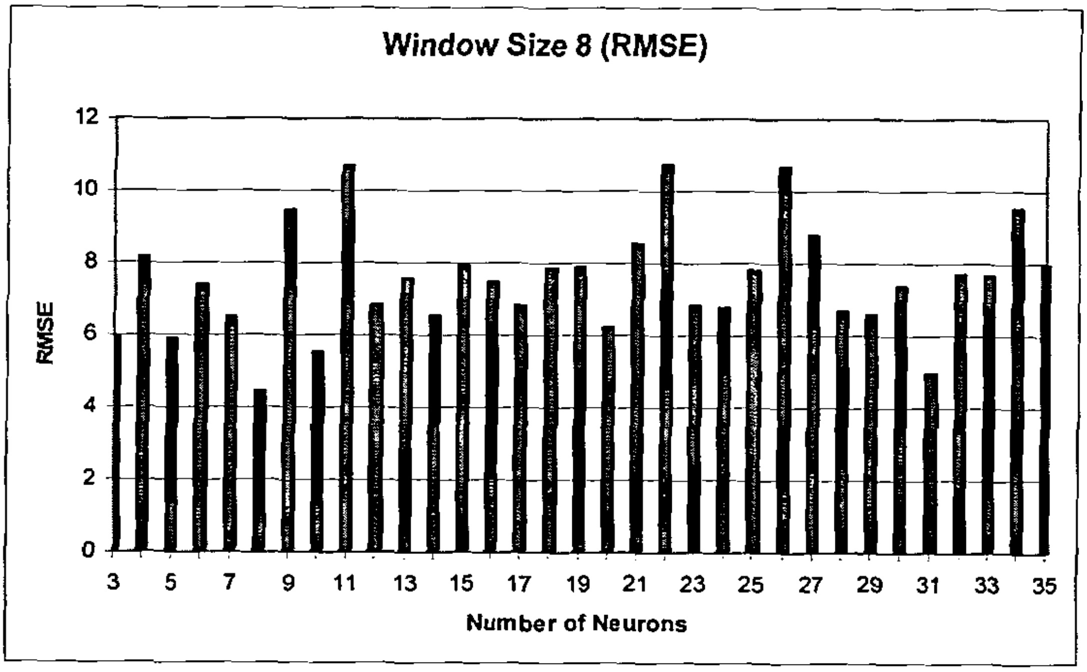

Figure A.5 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 8 and number of hidden neurons from 3 to 35.

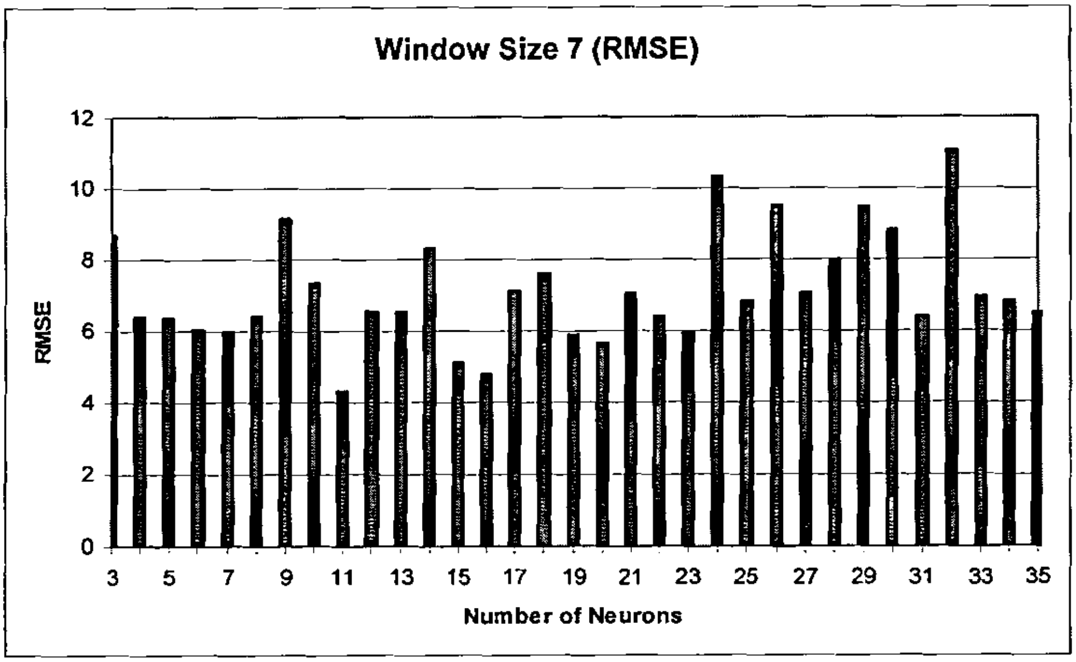

Figure A.6 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 7 and number of hidden neurons from 3 to 35.

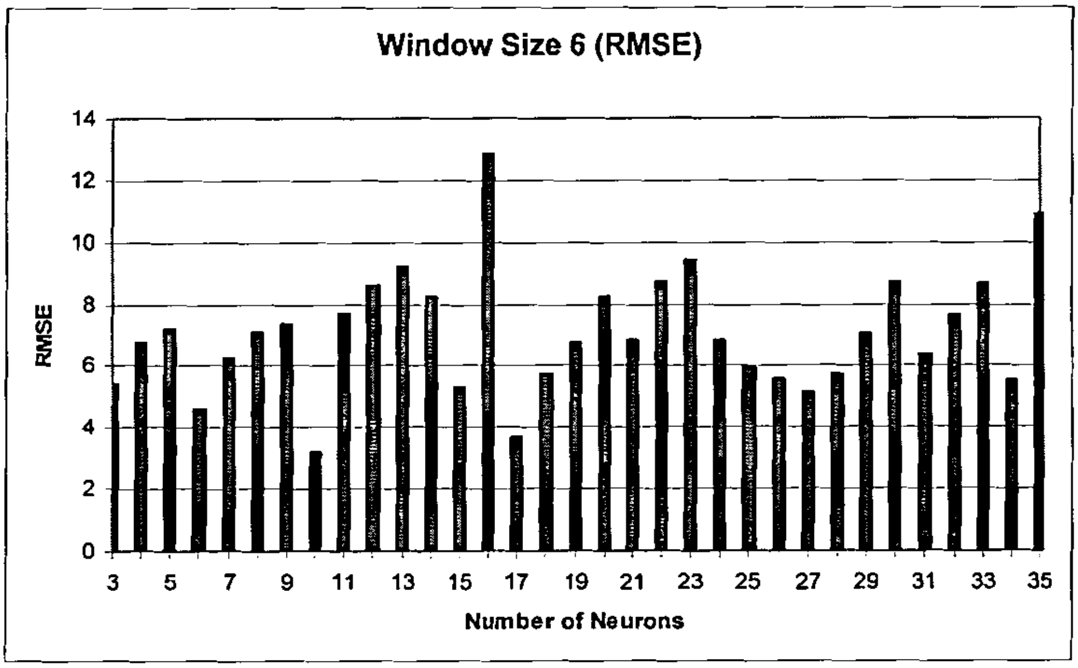

Figure A.7 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 6 and number of hidden neurons from 3 to 35.

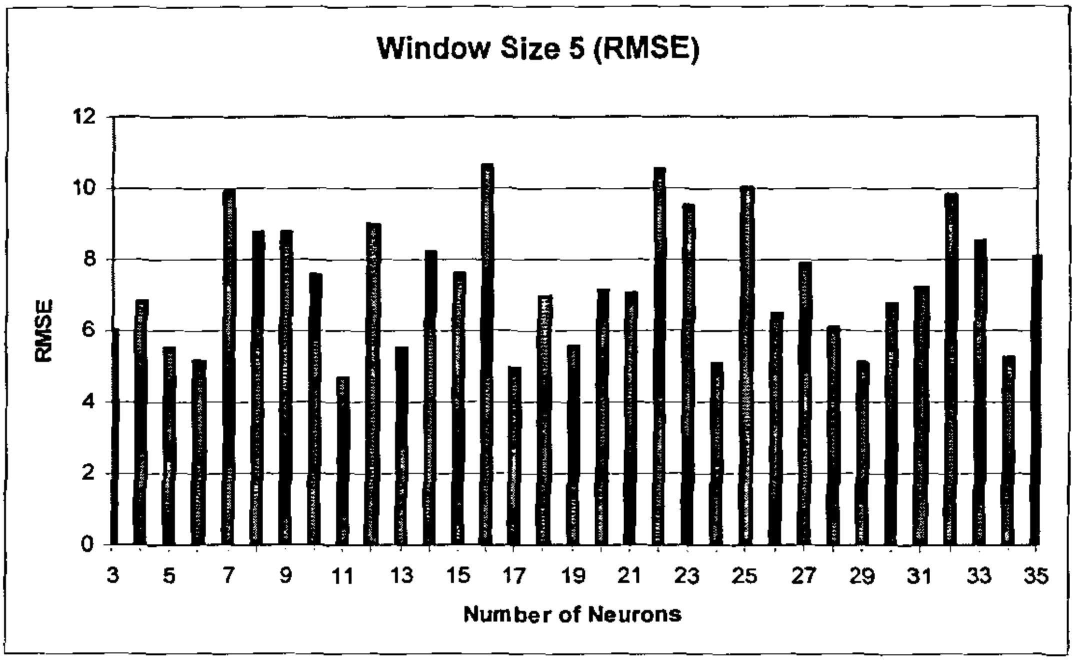

Figure A.8 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 5 and number of hidden neurons from 3 to 35.

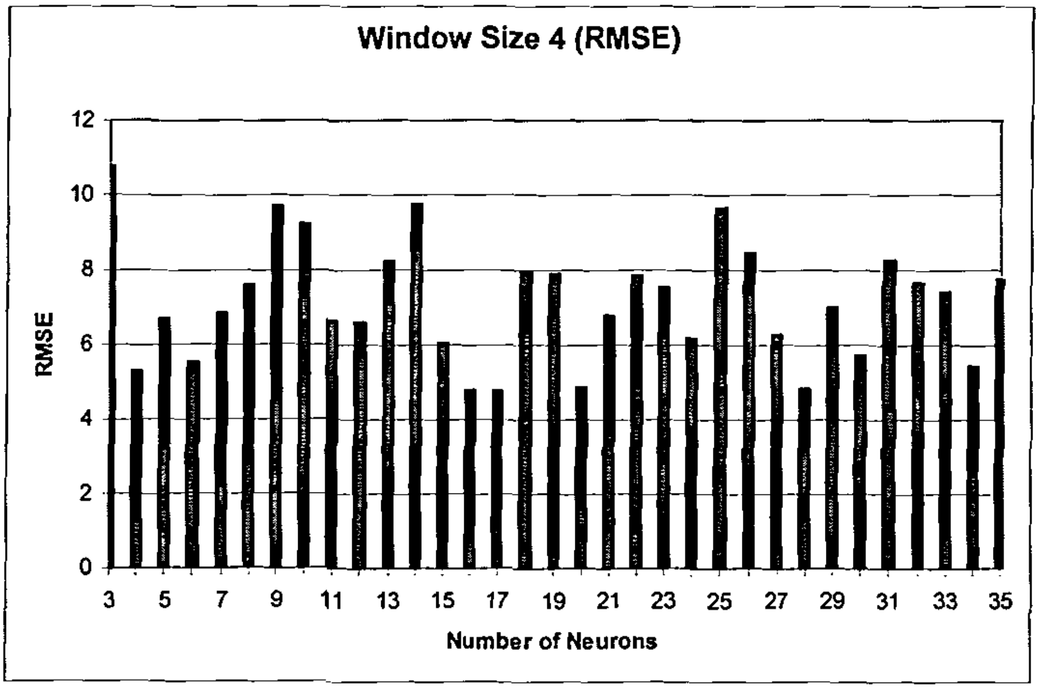

Figure A.9 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using Backpropagation with window size 4 and number of hidden neurons from 3 to 35.

A.2 RMSE using LVQ for Islamabad Summer Monsoon Rainfall Prediction

Figures below from Figure A.10 to Figure A.18 depict RMSE for Islamabad Monsoon Rainfall Prediction (2004-2008) using LVQ with window size (04 to 12) and hidden layer neurons (03 to 35).

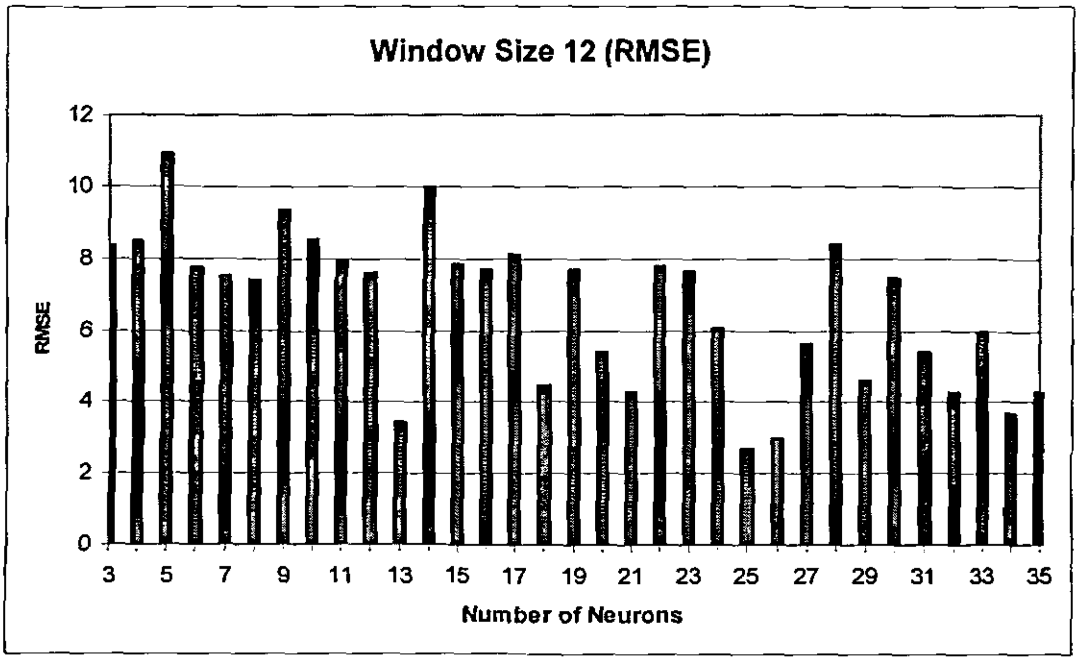

Figure A.10 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 12 and number of hidden layer neurons from 3 to 35.

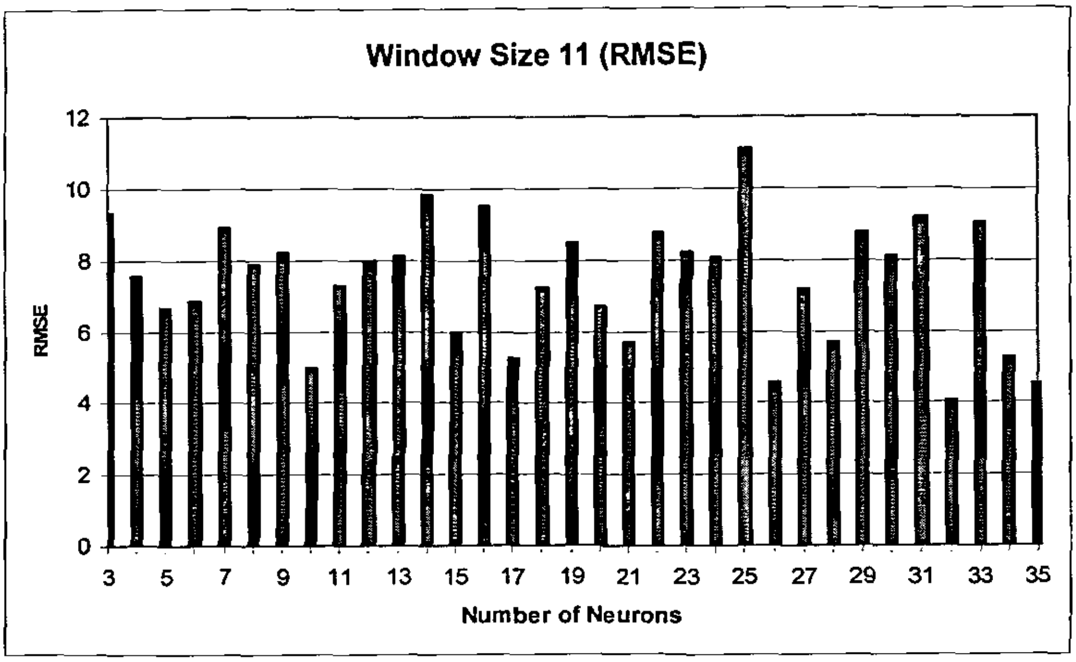

Figure A.11 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 11 and number of hidden neurons from 3 to 35.

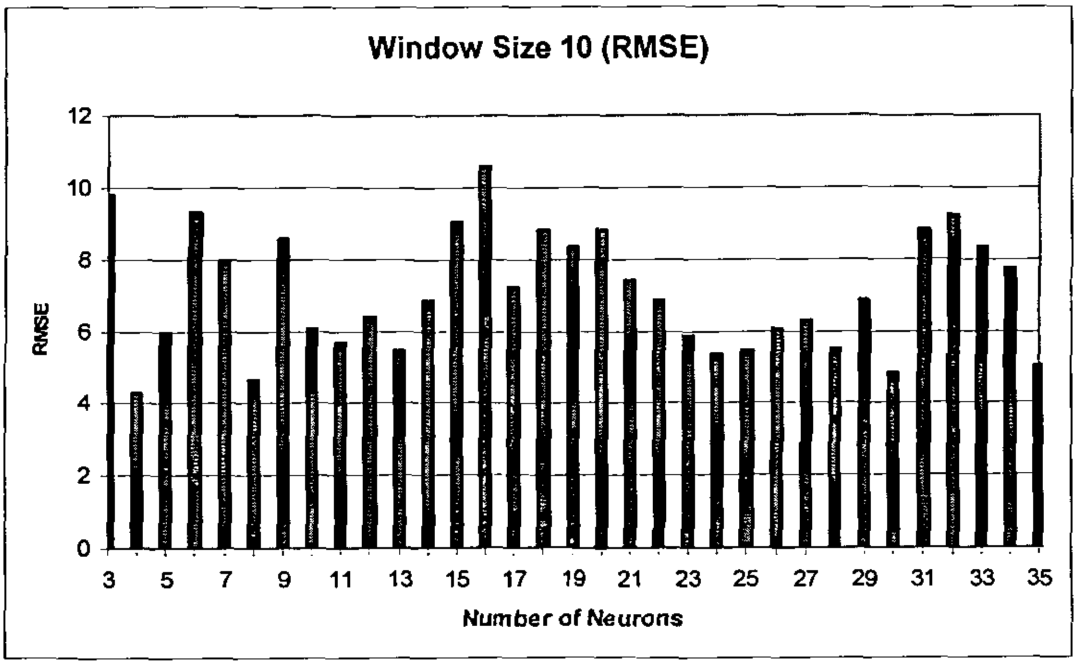

Figure A.12 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 10 and number of hidden neurons from 3 to 35.

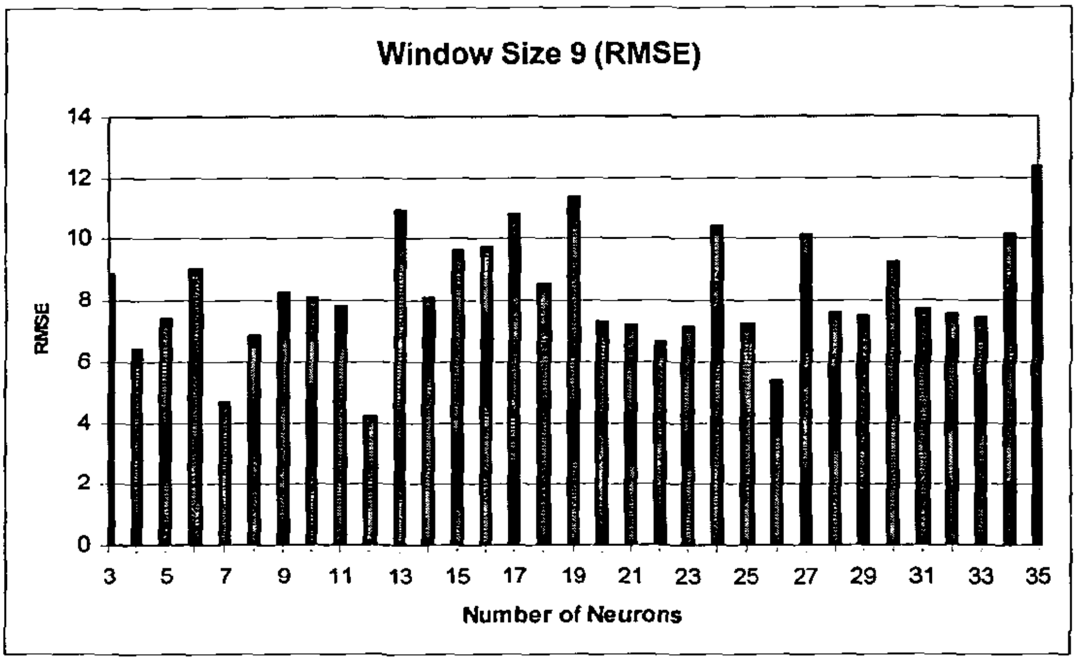

Figure A.13 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 9 and number of hidden neurons from 3 to 35.

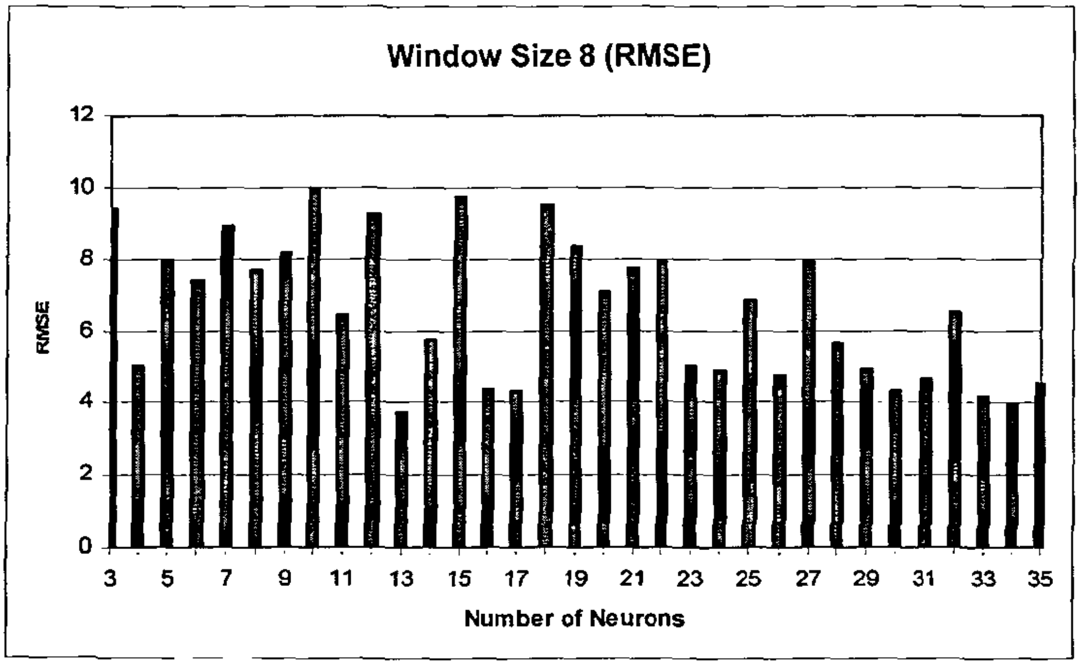

Figure A.14 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 8 and number of hidden neurons from 3 to 35.

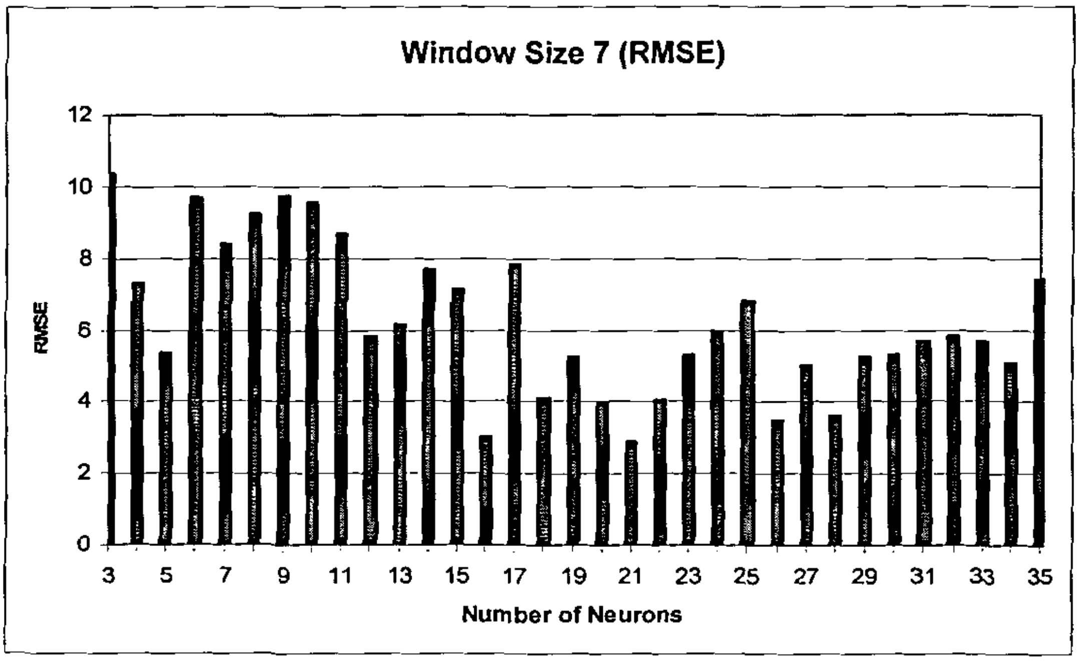

Figure A.15 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 7 and number of hidden neurons from 3 to 35.

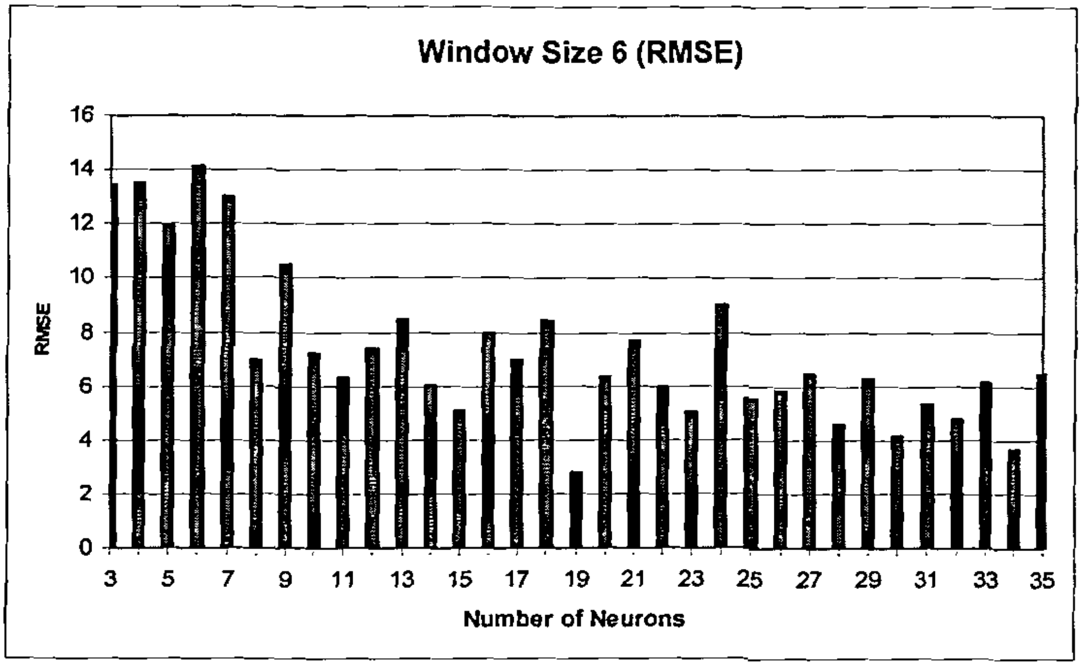

Figure A.16 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 6 and number of hidden neurons from 3 to 35.

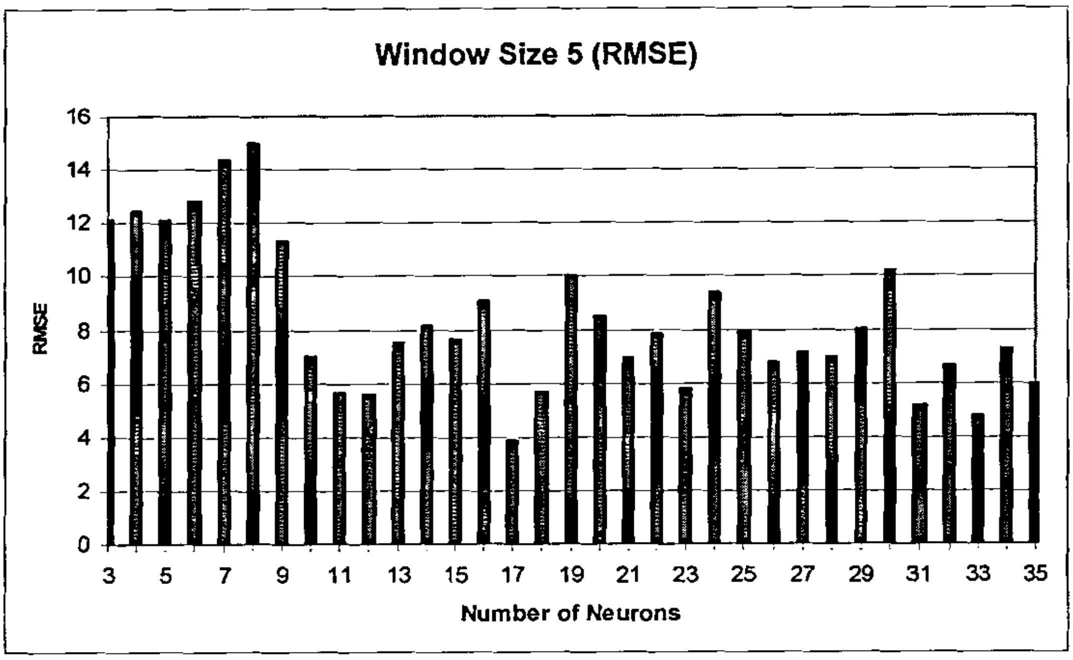

Figure A.17 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 5 and number of hidden neurons from 3 to 35.

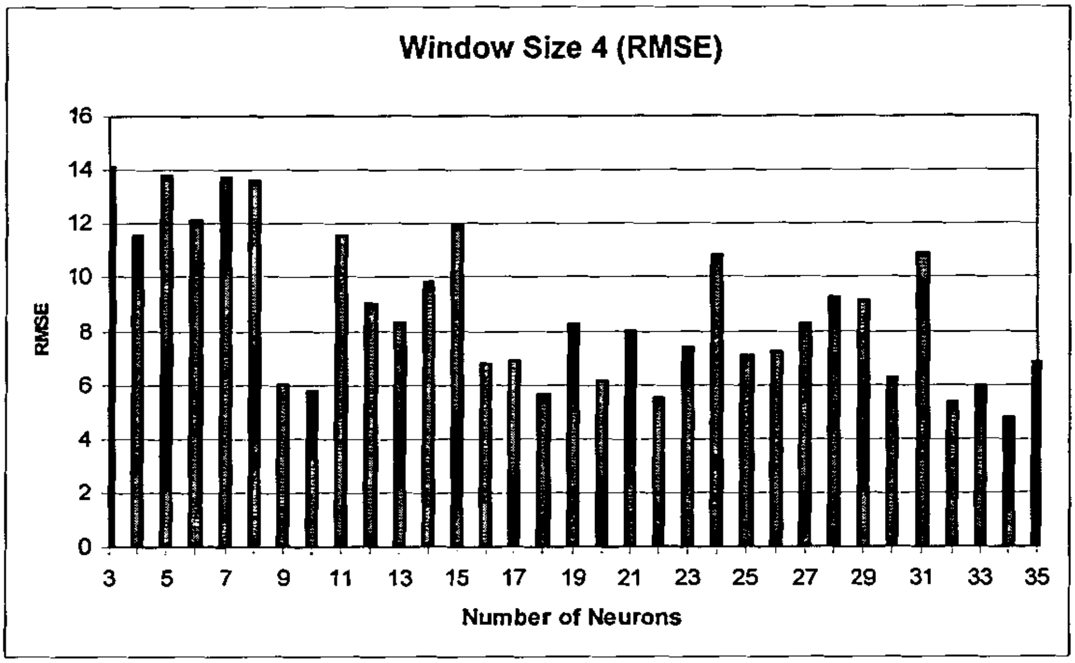

Figure A.18 RMSE (Class) on prediction of 5 years (2004-2008) Islamabad Summer Monsoon Rainfall using LVQ with window size 4 and number of hidden neurons from 3 to 35.

A.3 RMSE using Backpropagation for All Pakistan Summer Monsoon Rainfall Prediction

Figures below from Figure A.13 to Figure A.24 depict RMSE for Islamabad Monsoon Rainfall Prediction (2004-2008) using Backpropagation with window size (04 to 12) and hidden layer neurons (03 to 35).

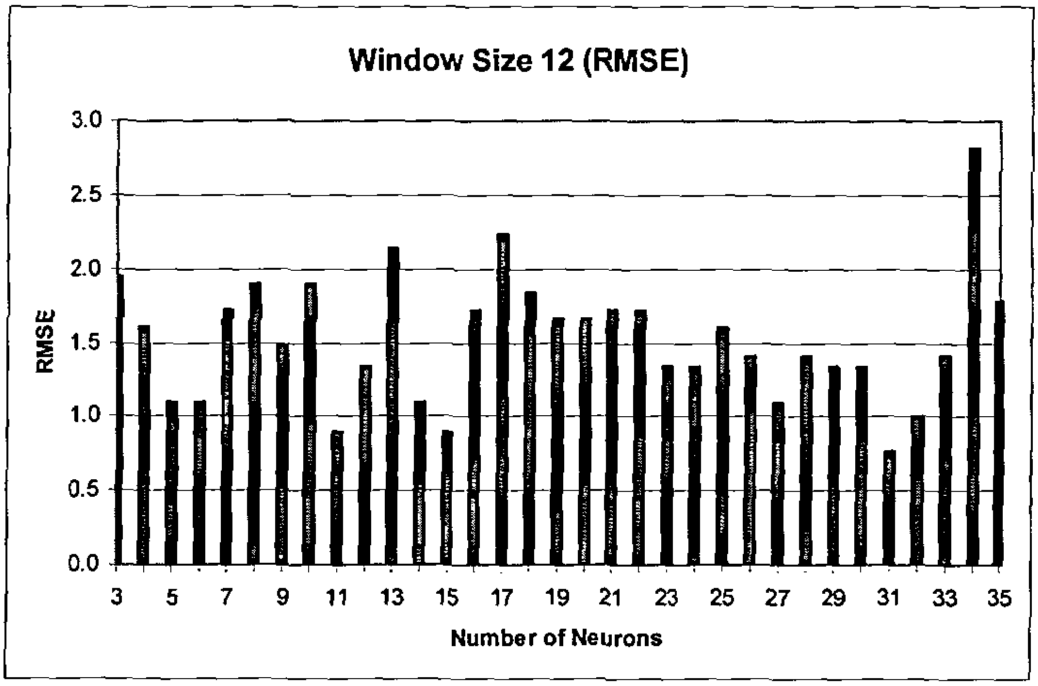

Figure A.19 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 12 and number of hidden neurons from 3 to 35.

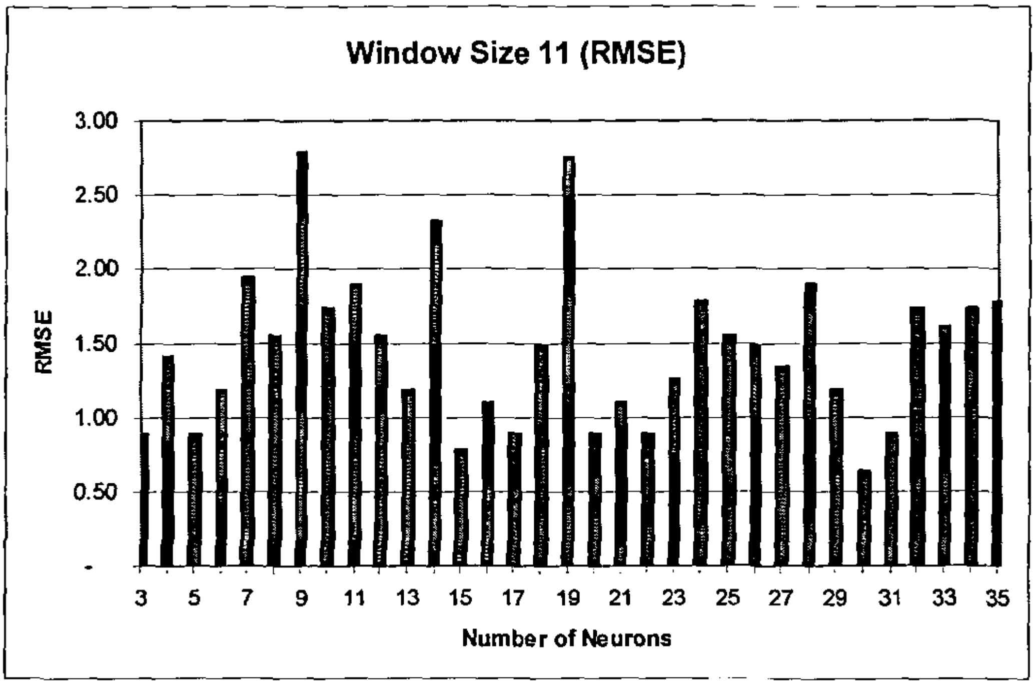

Figure A.20 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 11 and number of hidden neurons from 3 to 35.

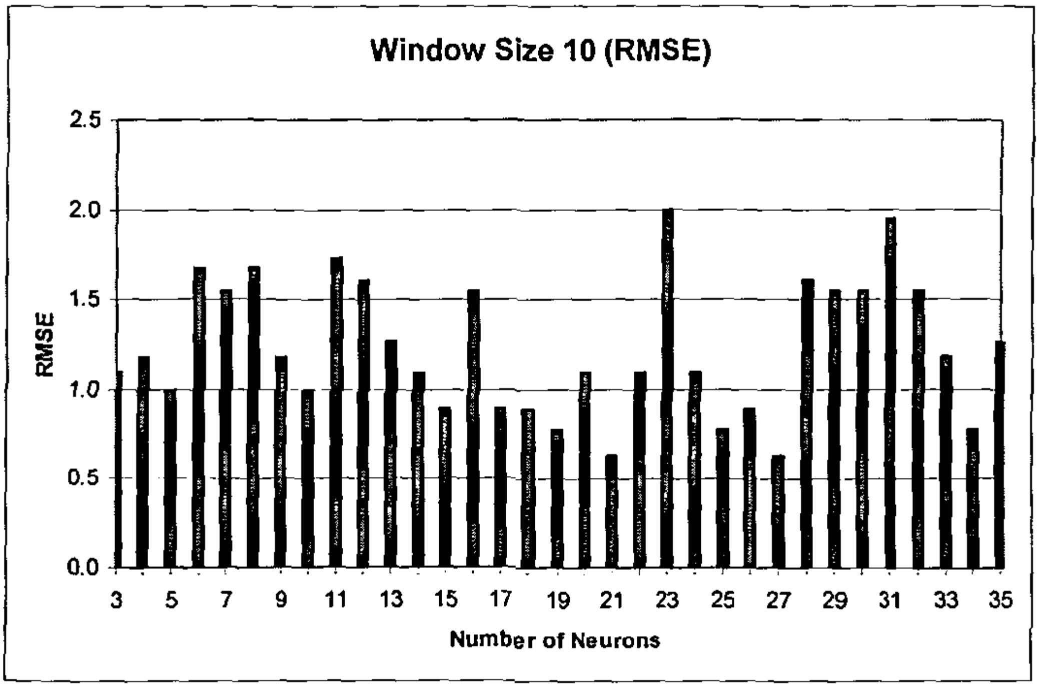

Figure A.21 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 10 and number of hidden neurons from 3 to 35.

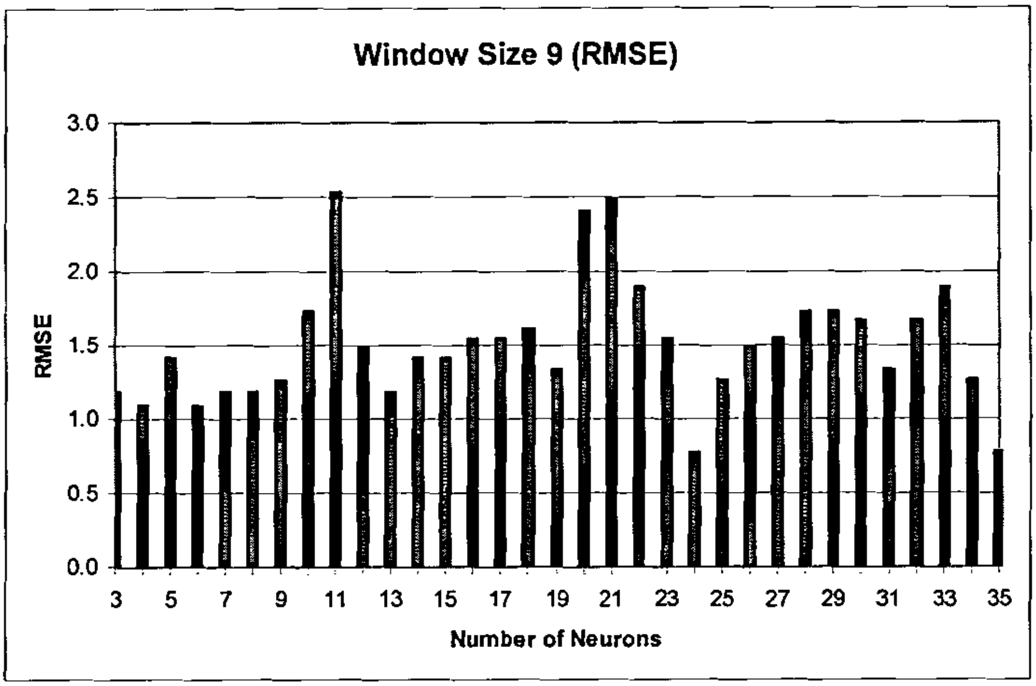

Figure A.22 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 9 and number of hidden neurons from 3 to 35.

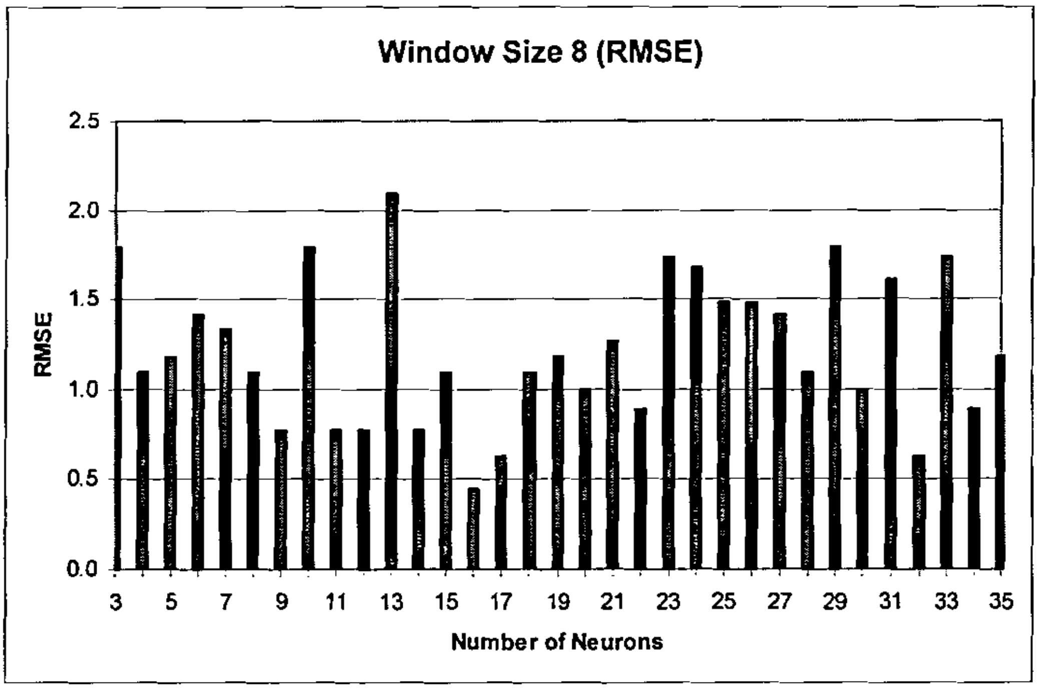

Figure A.23 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 8 and number of hidden neurons from 3 to 35.

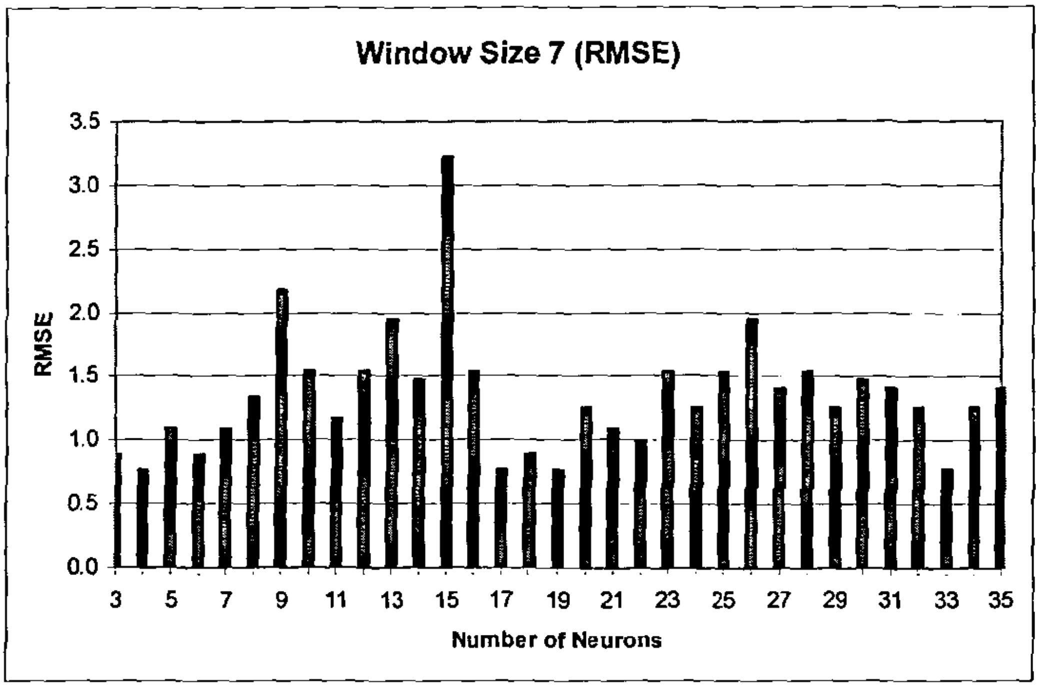

Figure A.24 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 7 and number of hidden neurons from 3 to 35.

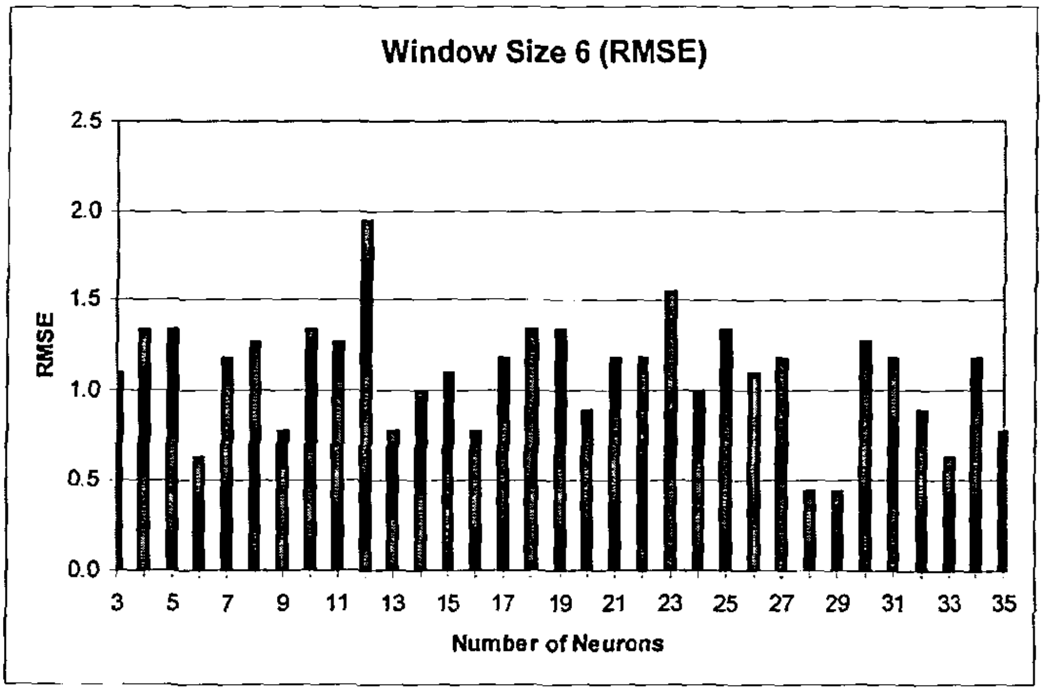

Figure A.25 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 6 and number of hidden neurons from 3 to 35.

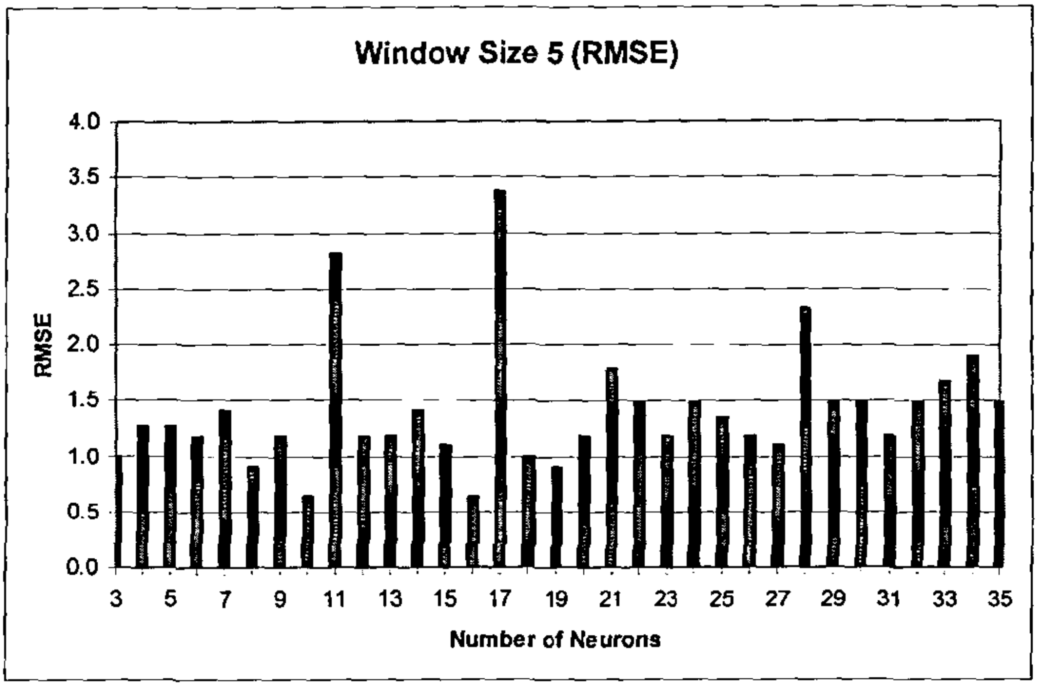

Figure A.26 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 5 and number of hidden neurons from 3 to 35.

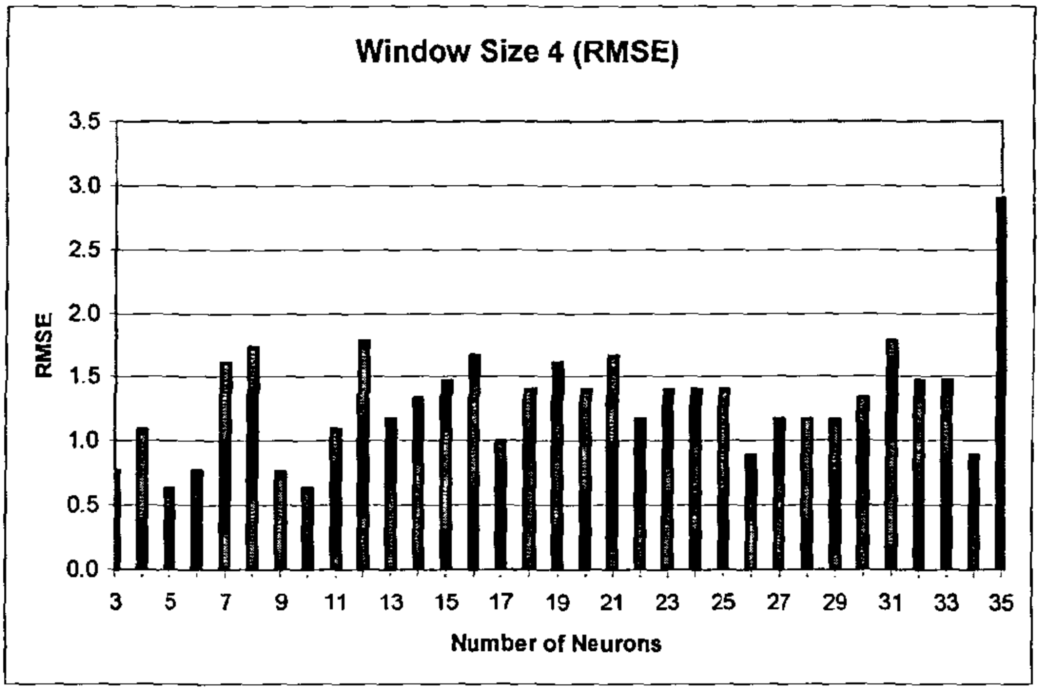

Figure A.27 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using Backpropagation with window size 4 and number of hidden neurons from 3 to 35.

A.4 RMSE using LVQ for All Pakistan Summer Monsoon Rainfall Prediction using

Figures below from Figure A.13 to Figure A.24 depict RMSE for Islamabad Monsoon Rainfall Prediction (2004-2008) using LVQ with window size (04 to 12) and hidden layer neurons (03 to 35).

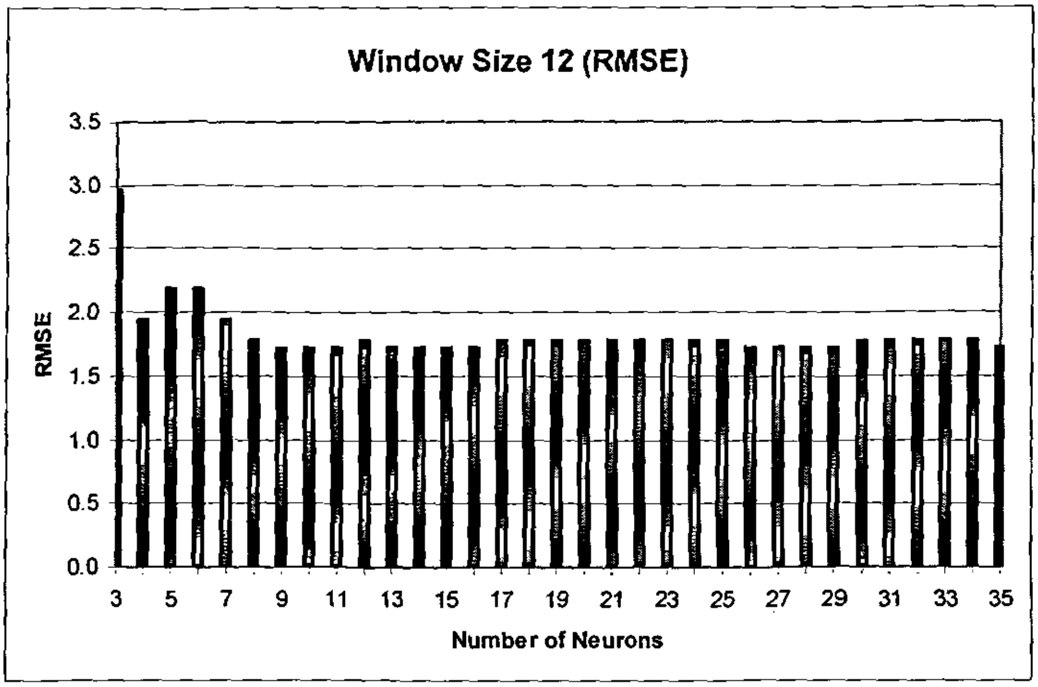

Figure A.28 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 12 and number of hidden neurons from 3 to 35.

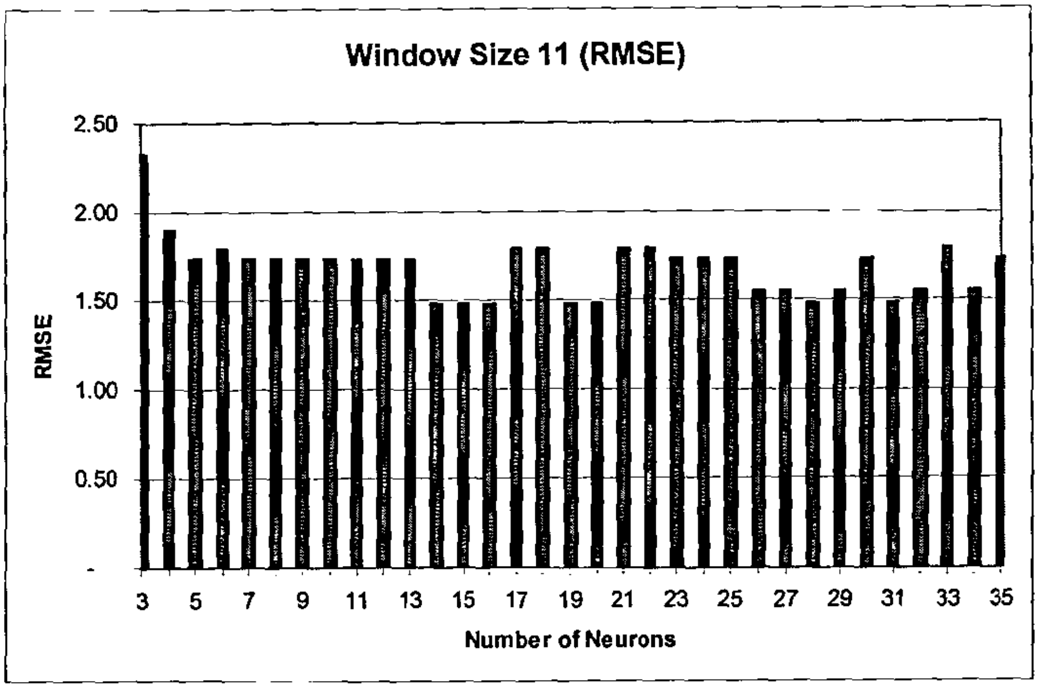

Figure A.29 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 11 and number of hidden neurons from 3 to 35.

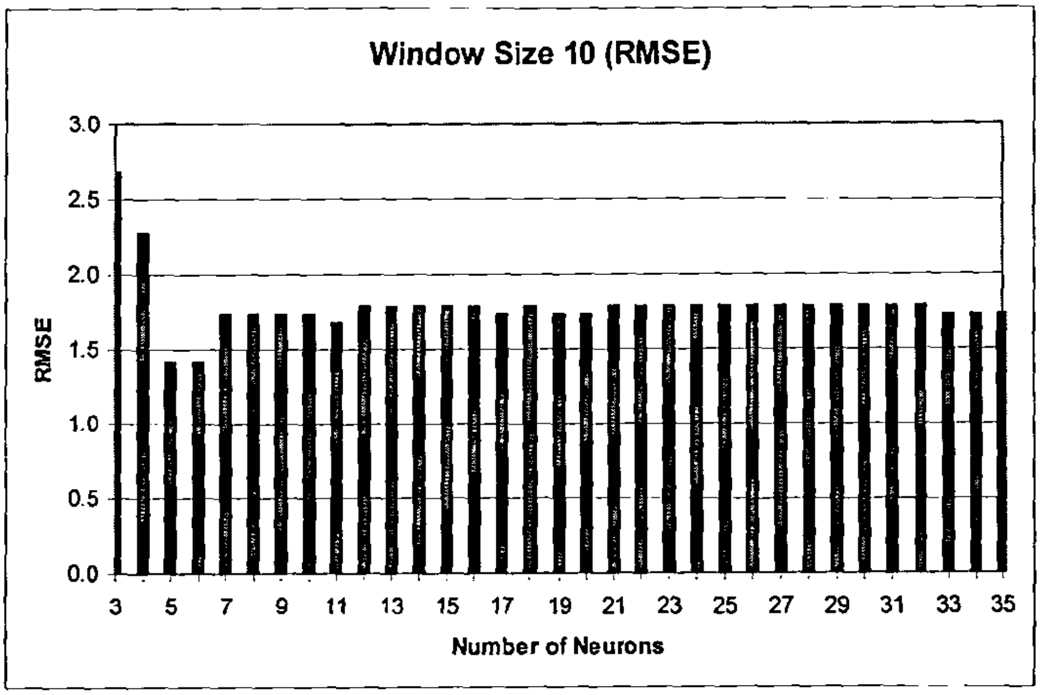

Figure A.30 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 10 and number of hidden neurons from 3 to 35.

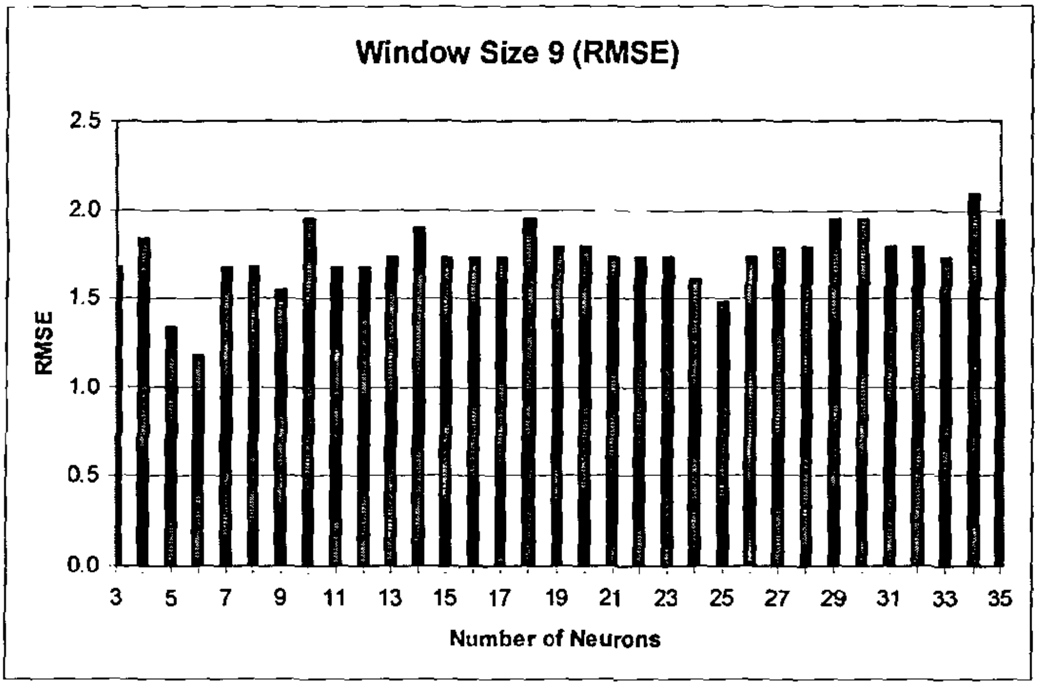

Figure A.31 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 9 and number of hidden neurons from 3 to 35.

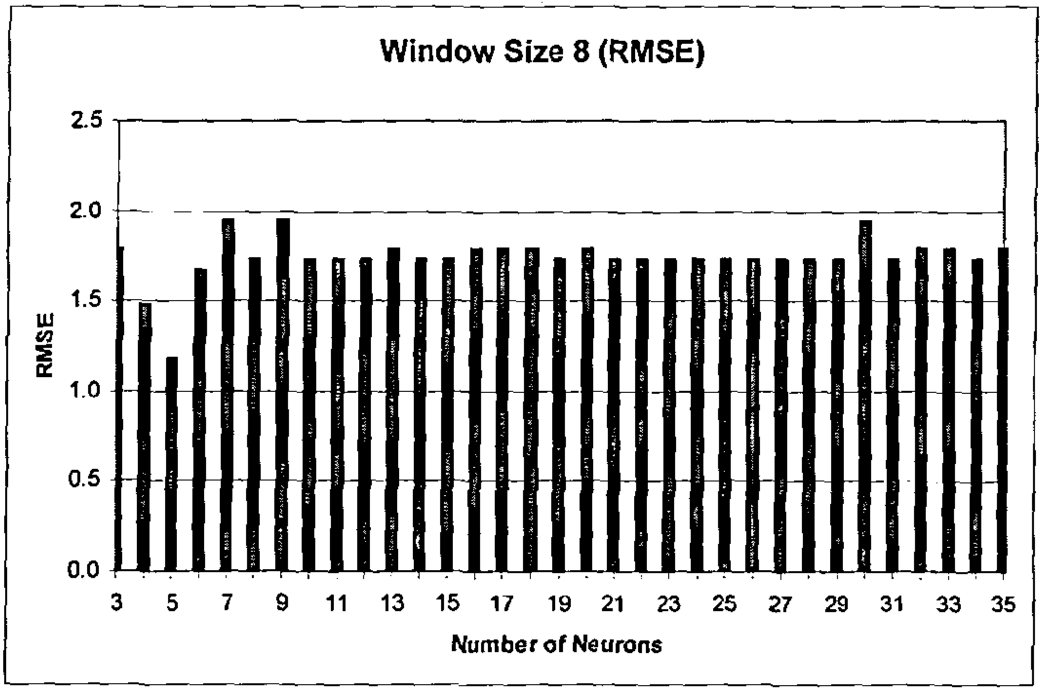

Figure A.32 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 8 and number of hidden neurons from 3 to 35.

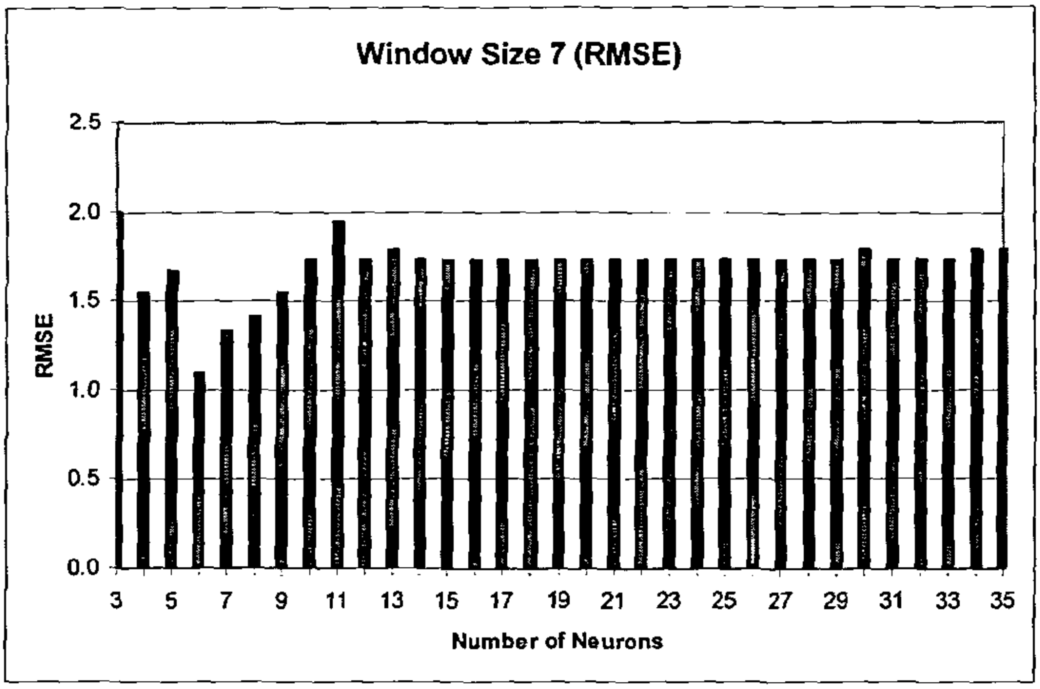

Figure A.33 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 7 and number of hidden neurons from 3 to 35.

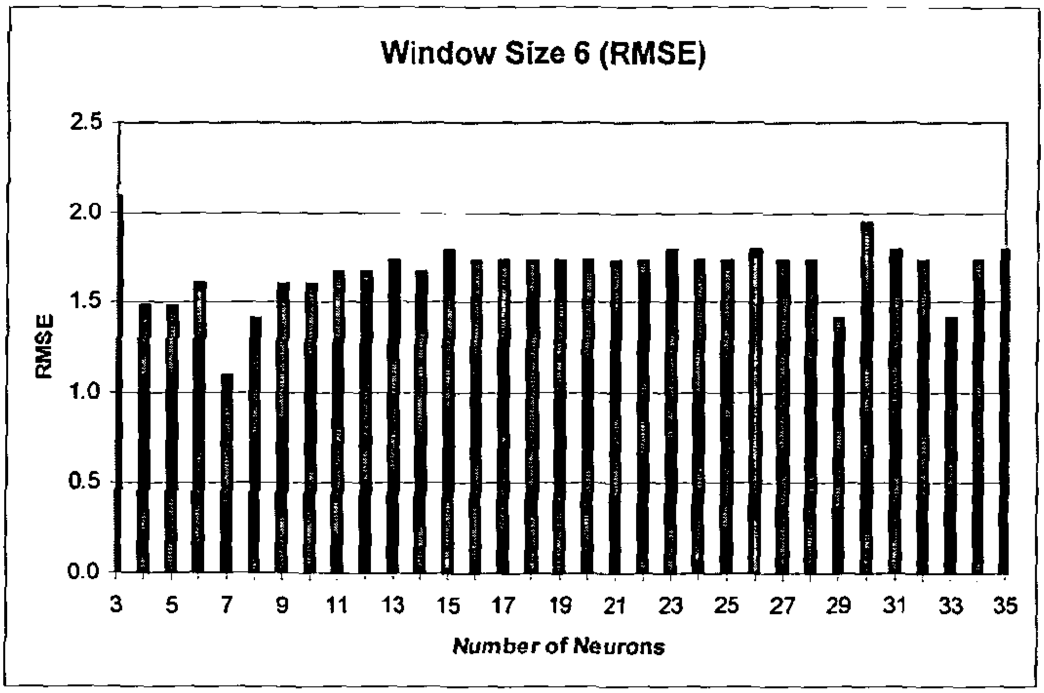

Figure A.34 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 6 and number of hidden neurons from 3 to 35.

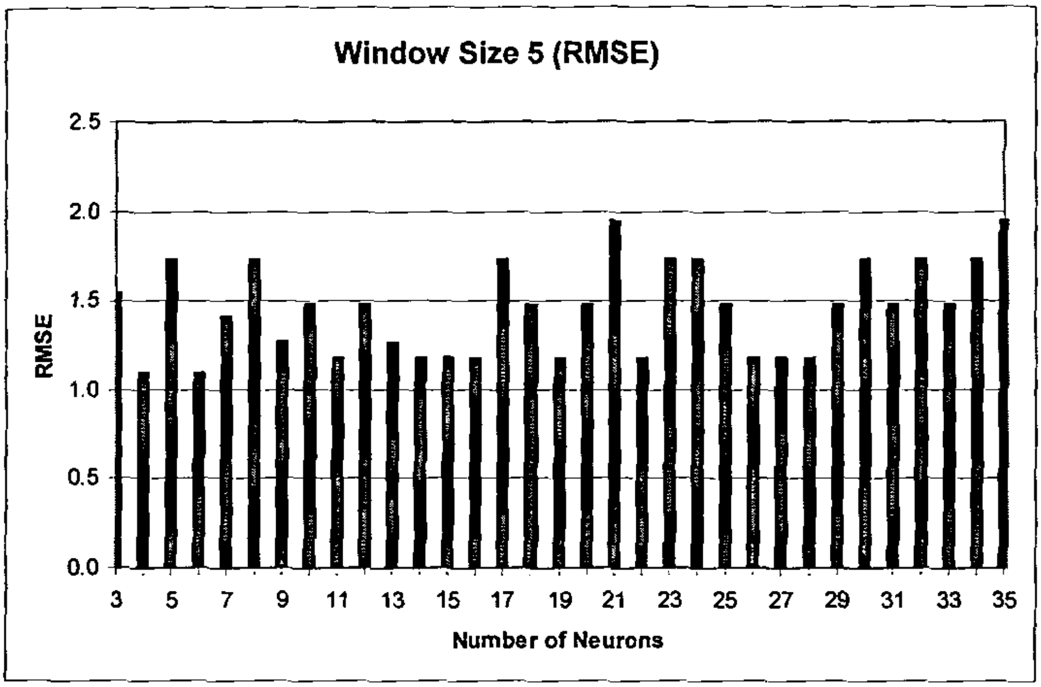

Figure A.35 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 5 and number of hidden neurons from 3 to 35.

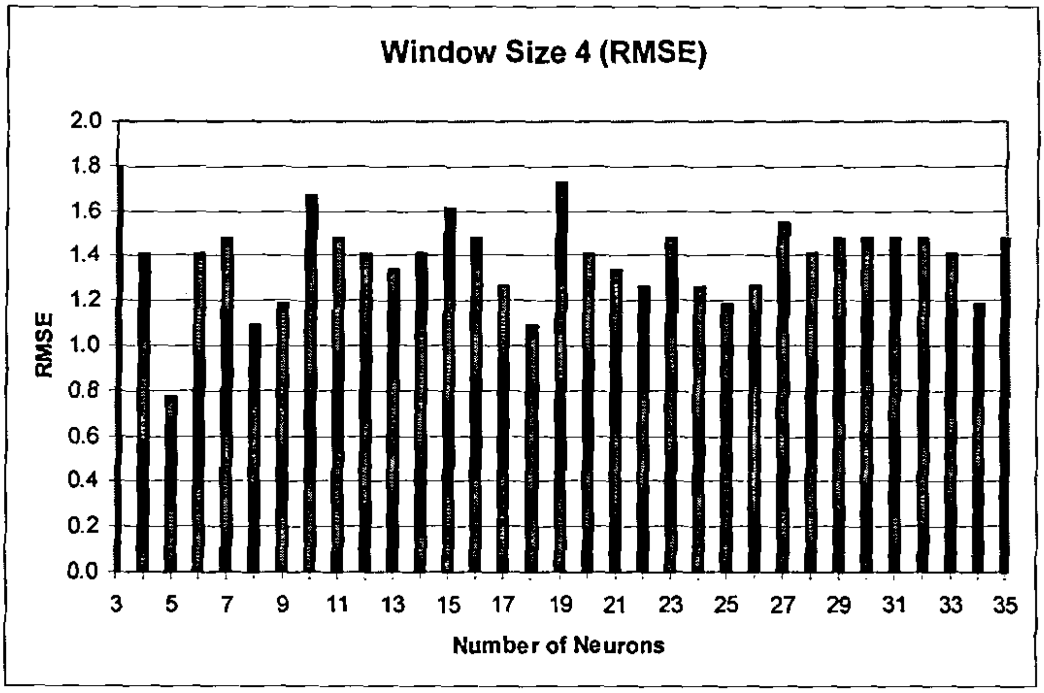

Figure A.36 RMSE (Class) on prediction of 5 years (2004-2008) All Pakistan Summer Monsoon Rainfall using LVQ with window size 4 and number of hidden neurons from 3 to 35.

References

References

[1] Q. Z. Chaudhary, "Analysis and seasonal prediction of Pakistan summer monsoon rainfall", *PhD thesis, Department of Meteorology and Oceanography, College of Science, University of Philippines, 1992.*

[2] "Annual Flood Report 2006", *Federal Flood Commission, Ministry of Water & Power, Pakistan, 2006.*

[3] G. T. Walker, "Correlations in seasonal variations of weather, II", *Memoirs of India Meteorological Department, 21, 22-45, 1910.*

[4] A. K. Banerjee, P. N. Sen and C. R. V. Raman, "On foreshadowing southwest monsoon rainfall over India with mid-tropospheric circulation anomaly of April", *Indian Journal of Meteorology, Hydrology and Geophysics, 29, 425–431, 1978.*

[5] E. C. Kung and T. Sharif, "A Long-range forecasting of the Indian summer monsoon onset and rainfall with upper air conditions", *Journal of Meteorological Society of Japan, 60, 672–681, 1982.*

[6] J. Shukla and D.A Mooley, "Empirical Prediction of The Summer Monsoon Rainfall Over India", *Monthly Weather Review, 115, 695-704, 1987.*

[7] V. Gowariker, V. Thapliyal, R.P. Sarker, G.S. Mandal, and D.R. Sikka, "Parametric and power regression models: New approach to long range forecasting of monsoon rainfall in India", *Mausam, 40, 115-122, 1989.*

[8] V. Gowariker, V. Thapliyal, S.M. Kulshrestha, G.S. Mandal, N. Sen Roy, and D.R. Sikka, "A power regression model for long range forecast of southwest monsoon rainfall over India", *Mausam, 42, 125-130, 1991.*

[9] B. Parthasarathy, K. Rupa Kumar and V. R. Deshpande, "Indian summer monsoon rainfall and 200mb meridional wind index: Application for long range prediction", *International Journal of Climatology, 11, 165–176, 1991.*

[10] M. Rajeevan, D. S. Pai, S. K. Dikshit and R. R. Kelkar, "IMD's new operational models for long-range forecast of southwest monsoon rainfall over India and their verification for 2003", *Current Science, Vol. 86, No. 3, February 2004.*

[11] A. A. Munot and K. Krishna Kumar, "Long range prediction of Indian summer monsoon rainfall", *Journal of Earth System Science 116, No. 1, pp. 73–79, February 2007.*

[12] S. Gadgil, M. Rajeevan and R. Nanjundiah, " Monsoon prediction – Why yet another failure?", *Current Science, Vol. 88, No. 9, May 2005.*

[13] A. Karori, "Downscaling NCC CGCM output for seasonal precipitation prediction over Islamabad – Pakistan", *Pakistan Journal of Meteorology, Vol. 4. Issue 8, January 2008.*

[14] S. Mitra and T. Acharya, "Data Mining: Multimedia, Soft Computing, and Bioinformatics", *John Wiley and Sons, 2003.*

[15] P. Guhathakurta, "Long lead monsoon rainfall prediction for meteorological sub-divisions of India using deterministic artificial neural network model", *Meteorology and Atmospheric Physics, Vol. 101, No. 1, pp. 93-108, 2008.*

[16] S. Karmakar, M.K. Kowar and P. Guhathakurta, "Development of an 8-Parameter Probabilistic Artificial Neural Network Model for Long-Range Monsoon Rainfall Pattern Recognition over the Smaller Scale Geographical Region – District", *First International Conference on Emerging Trends in Engineering and Technology (ICETET '08), 2008.*

[17] S. Karmakar, M.K. Kowar and P. Guhathakurta, "Long-Range Monsoon Rainfall Pattern Recognition and Prediction for the Subdivision 'EPMB' Chhattisgarh Using Deterministic and Probabilistic Neural Network", *Seventh International Conference on Advances in Pattern Recognition (ICAPR '09), 2009.*

[18] P. Guhathakurta, "Long range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network", *Current Science Vol. 90, No. 6, 2006.*

[19] H.D. Navone and H.A. Ceccatto, "Predicting Indian monsoon rainfall: a neural network approach", *Climate Dynamics, Vol. 10, 1994.*

[20] P. Goswami and P. Kumar, "Experimental annual forecast of all India mean summer monsoon rainfall for 1997 using a neural network model", *Current Science Vol. 72, 1997.*

[21] S. Banik, F.H. Chanchary, K. Khan, R.A. Rouf and M. Anwer, "Neural network and genetic algorithm approaches for forecasting Bangladeshi monsoon rainfall", *11th International Conference on Computer and Information Technology(ICCIT 2008), 2008.*

[22] J. McCullagh, K. Bluff and T. Hendtlass, "Evolving Expert Neural Networks for Meteorological Rainfall Estimations", *Proceedings of the 1999 International Conference on Neural Information Processing and Intelligent Information Systems (ICONIP'99), 2, 585-590, 1999.*

[23] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou and K. Menagias, "Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values", *International Journal of Mathematical, Physical and Engineering Sciences, Vol. 1, No. 1, pp.16-20, 2007.*

[24] N. Dong-Xiao, G. Zhi-Hong, X. Mian and W. Hui-Qing, "Approach to Daily Load Forecast of VSNN Based on Data Mining", *2nd IEEE Conference on Industrial Electronics and Applications (ICIEA 2007), 2007.*

[25] M. Xing and W. Sun, "Data Mining Based Neural Network Model for Load Forecasting", *Proceedings of 2005 International Conference on Machine Learning and Cybernetics, 2005.*

[26] R. Li, J. H. Li and H. M. Li, "The Short –Term Electric Load Forecasting Grid Model Based on MDRBR Algorithm", *IEEE Power Engineering Society General Meeting, 2006.*

[27] N. Dong-xiao and W. Yong-li, "Study of the SMO Algorithm Based on Data Mining in Short-Term Power Load Forecasting Model", *7th World Congress on Intelligent Control and Automation, (WCICA 2008), 2008.*

[28] P. Mandal, T. Senjyu, K. Uezato and T. Funabashi, "Several-Hours-Ahead Electricity Price and Load Forecasting Using Neural Networks", *IEEE Power Engineering Society General Meeting, 2005.*

[29] P. Mandal, T. Senjyu and T. Funabashi, "Neural Network Models to Predict Short-Term Electricity Prices and Loads", *IEEE International Conference on Industrial Technology, (ICIT 2005), 2005.*

[30] N. Dong-xiao, W. Yong-li "Support Vector Machines Based on Data Mining Technology in Power Load forecasting", *International Conference on Wireless Communications, Networking and Mobile Computing (WiCom 2007), 2007.*

[31] D. M. A. Hussain, A.Q. K. Rajput, B. S. Chowdhry and Q. Gee, "Seasonal to Inter-annual Climate Prediction Using Data Mining KNN Technique", *International Multi Topic Conference on Wireless Networks, Information Processing and Systems, IMTIC 2008, Jamshoro, Pakistan.*

[32] J. Han and M. Kamber, "Data mining: concepts and techniques", *Edition: 2, Morgan Kaufmann, 2006.*

[33] C. Goodess, J. Palutikof and M. Agnew, "Climate Change Scenario for the Mediterranean: A Basis for Regional Impact Assessment", *International Workshop "Climate Change and Mediterranean Coastal Systems: Regional Scenarios and Vulnerability Assessment", December 9-10, 1999.*