International Journal of Scientific and Innovative Mathematical Research (IJSIMR)
Volume 3, Issue 6, June 2015, PP 22-28
ISSN 2347-307X (Print) & ISSN 2347-3142 (Online)
www.arcjournals.org

Authorship Attribution of Tamil Articles using Artificial Neural Network

R. Lakshmi Priya
Department of Statistics
Dr. Ambedkar Govt. Arts College
Vysarpadi, Chennai, INDIA.

G. Manimannan
Department of Statistics
Madras Christian College
Chennai, INDIA

Abstract: *This paper attempts to build an Artificial Neural Network that can classify distinguish authorship problem of three contemporary Tamil scholars of the similar period, namely Mahakavi Bharathiar (MB), Subramaniya Iyer (SI), and T. V. Kalyanasundaranar (TVK). The three trendy scholars stated above had written more number of articles on India's Freedom Movement during the pre-independence period and published in the magazine called, India in the year 1906 is considered for this study. Artificial Neural Networks have been used newly as a modern tool and attracted broad research in various disciplines including Authorship Attribution. Application of Neural Network models has increased considerably in areas of pattern recognition and classification problems in the field of Authorship over the last decade. A set of variables such as function words is made use of classification purposes.*

Keywords: *Authorship Attribution, Tamil Articles, Classification, Artificial Neural Network.*

1. INTRODUCTION

The discipline connected with *authorship* paved the path to the problem of authorship attribution. *Authorship* is the study of the quantifiable of human language, or the statistical analysis of literary style (Holmes, 1995; Holmes and Forsyth, 1995). This involves attempting to formally capture the creative, unconscious elements of language particular to individual writers and speakers. Although researchers have studied writing for centuries, the discipline of stylometry is fairly recent, and while its origins date back to the late 19th century, the field as it is now began with work on the *Federalist Papers* in 1968 (Mosteller and Wallace, 1968).

Stylometry mainly concerns itself with authorship attribution studies, although chronological studies on the dating of work within the corpus of an author have also investigated. Writing in a forensic background, Bailey (1979) proposed three rules to define the situation necessary for authorship attribution:

1. The number of putative authors should constitute a well- defined set.
2. The lengths of the writings should be sufficient to reflect the linguistic behavior of the author of the disputed text and also those of the candidates.
3. The texts used for comparison should be commensurate with the disputed writing.

A computational stylistic study of doubtful authorship should involve comparisons of the disputed text with works by each of the possible candidate authors using suitable statistical tools on quantifiable features of the texts-features which reflect the style of the writing as defined above.

One modern addition to the tools available for the computational stylometry is that of the Artificial Neural Network (ANN). These are the computational methods closely based on the concept of biological neuron, the idea being that simple, trained processing elements will result in much more difficult behavior when used in combination.

In recent years, many scholars have successfully demonstrated that this technique of machine learning field can be applied to authorship attribution. Merriam and Mathews (1993, 1994) have trained a multi layer perception network to distinguish the works of Shakespeare and Marlowe. Tweedie *et al.* (1996) have provided a useful review of the applications of ANNs in the area of computational stylometry

and have used this machine-learning package for the reanalysis of the *Federalist Papers*. Kjell (1994) have taken up authorship study using letter-pair frequency features with neural network classification. The present study attempts to use the Neural Network Pattern Recognition (NNPR) as one of the suitable neural network classification tool.

2. NEURAL NETWORKS

An Artificial Neural Network (ANN) is a mathematical model represented by interlinked simple computational elements, called neurons that could compute, learn, remember and optimize the way a human brain works (Bishop, 2003; Heykin, 2001; Wasserman, 1989, 1993). The neurons in ANN are called nodes. The interconnected nodes (neurons) are arranged into several layers namely, the input, intermediate (hidden) and the output layers. Depending on the signals (data) transmitted by various nodes, a set of outputs is computed by the nodes that received the signals from the other nodes.

Initially the network should be subjected to learning process. The network must learn decision surfaces from a set of training patterns so that these training patterns are classified correctly (Gose, Johnsonbaugh and Jost, 1997). After training, the network must also be able to generalize, that is, correctly classify test patterns it has never seen before. Usually any neural network should be such that it has the ability to learn well and also to generalize well. The general procedure is to have the network learn the appropriate weights from a representative set of training data. In all but the simplest cases, however, direct computation of the weights is difficult. Instead, learning starts off with *random initial weights* and adjusts them in successive iterations, until the required outputs are produced.

The supervised and the unsupervised learning methodologies are adapted by the ANN. In supervised learning, the objective is to predict one or more output variables from one or more input variables (Bishop, 2003; Ripley, 1996). In the unsupervised learning, there are no target variables. The network trains itself to extract the features from the input variables using which the input variable themselves can be predicted.

Tweedie e*t. al.* (1996) discussed the application of neural networks in stylometry and their usefulness for a number of reasons:

1. Neural networks can learn from the data themselves. Implementing a rule-based system in linguistic computing may become complex as the number of distinguishing variables increases and even the most complex rules may still not be good enough to completely characterize the training data. In essence, neural networks are more adaptive.
2. Neural networks can generalize. This ability is particularly required in the literary field, as only limited data may be available.
3. Neural networks can capture non-linear interactions between input variables.
4. Neural networks are capable of fault tolerance. Hence a particular work, which is not in line with the usual writing style of an author, will not affect the network to a considerable extent. Thus neural networks appear to promise much for the field of stylometry. Their application would appear to be worthy of investigation.

The pioneering work in the application of neural nets in Stylometry was undertaken by Merriam and Mathews (1993). In their paper, a very small set of function word frequencies is used as input to a multiplayer perceptron (a neural net having a hidden layer) to examine four plays that have been attributed both to Shakespeare and John Fletcher. In the present context an attempted is made to extend the use of the concepts of Neural Network Pattern Recognition tool for the authorship attribution problem (Chandrasekaran R. and Manimannan G, 2008).

3. NEURAL NETWORK PATTERN RECOGNITION USING MATLAB

The main objective of pattern recognition algorithm is to classify authorship problem interests into the correct class among the three groups of authors of MB, TVK and SI. These authors of interest are function words, variables and the authorship objective is to classify them into three different tasks achieved. For this reason, artificial neural networks were selected. In this research paper the following algorithm is used for performing classification of three authors, namely, Mahakavi Bharathiar (MB), T. V. Kalyanasundaranar (TVK) and Subrmaniya Iyer (SI). One familiar classification method of Neural Network is the scaled conjugate gradient algorithm to use authorship classification.

3.1 Algorithm

Scaled Conjugate Gradient (SCG) algorithm- It's a second order conjugate algorithm which helps in minimizing the goal functions of several parameters. It uses a step size scaling mechanism which makes it faster than other second order algorithms by avoiding line search per learning iteration (Martin Riedmiller and Heinrich Braun, 1993).

In this context the following steps are made use of MATLAB for the present research:

1. Initially open the Neural Network Pattern Recognition tool box to classify inputs into a set of target categories.
2. The input data is $n*m$ matrix, where n representing samples size and m representing target data matrices of different class.
3. For validation and test, randomly divide up n samples. The selection of samples is by default, $n=70$ percentage of sample for training $n=15$ samples for validation and remaining $n=15$ percentage of samples for testing.
4. To fix the network size based on input layer a , hidden layer b and output layer c , output d , where $a = number\ \ of\ \ input$,
   $b = number\ \ of\ \ hidden\ layer\ (by\ defaut\ \ size\ or\ mannual)$,
   $c = number\ \ of\ output\ \ layer$,
   $d = number\ \ of\ output\ \ class$.
5. Train the network to classify the input data matrix according to the targets. Train the networks until your goal. Sometimes, training multiple times of data matrix will generate different results due to different initial conditions and sampling.
6. The result is displayed in the following order, Training, Validation and Testing sample in the second column, Training, Validation and Testing MSE in the third column, Training, Validation and Testing percentage of error in the last column.
7. Finally, the neural network training window consists of Network diagram, algorithms, progress, plots and plots interval.

4. DATABASE

The present study deals with the literary works of three contemporary Tamil scholars, namely, Mahakavi Bharathi (MB), T. V.Kalyanasundaranar (TVK), and Subramaniya Iyer (SI). In the Pre–Independence period, these three scholars have written number of articles on India's Freedom Movement in the magazine called *India*. Initially, all the three scholars have written articles by attributing their names. The oppressive attitude of the then British Regime made all the three writers to write articles on the same topic anonymously in the same magazine. For this quantitative attribution study, all attributed articles of these three scholars written on India's Freedom Movement in the year 1906 are considered. Our study is based on thirty five blocks of MB, thirty eight blocks of TVK and thirty one block of SI. Each block contains ten sentences. To quantify the writing style of each sentence the researchers used function words (twenty four) variables. The samples of lists of parameters of this study with their Tamil meanings are given in *Table 1*.

Table 1. *Lists of Sample of Function words of this Study with Abbreviations*

Function Words	Translation	Function Words	Translation
மேலும்	Also	குறிக்கப்படாத	Unmarked
படி/பிரகாரம்	As	கொண்டு	With
என்றால்	If	செய்ய	To
ஆக	For	இல்	In
மிகவும்	As	கொண்டு	With
போல	Very much	எனும்	The
என்று	Alike	உடைய	of
பேரில்	That	பற்றி	About
இருந்து	To	குறைந்தது	At least
கூட	From	உள்ளே	Inside
	Also		

A chi-square analysis of the thirty five blocks of MB establishes that these articles do not differ from one another in terms of the frequency distribution of occurrence of these stylistic features. Similar results were obtained in the case of other two scholars (Manimannan and Bagavandas, 2001). Thus, each article is converted as a raw data matrix and these raw data matrices form the basis for this data description. Hence the thirty five blocks of MB consist of three hundred and fifty sentences, the thirty one blocks of SI consist of three hundred and ten sentences and thirty eight blocks of TVK consist of three hundred and eighty sentences.

5. RESULTS AND DISCUSSION

Thirty five blocks written by Mahakavi Bharathiar (MB), thirty one blocks by Subramaniya Iyer (SI) and thirty eight blocks by T. V. Kalyanasundaram (TVK) have been considered for the analysis represented by the averages of the sampled sentences. The 1040 data points representing the means of the corresponding sample writings of three different authors are used. The Data matrix (P) consists of 24 function words variables, normalized, computed from the chosen articles. The Target matrix (T) has three rows, one row for each author, each row consisting of zeroes and ones. This matrix element t_{ij} is defined as

$$t_{ij} = 1 \ \ if \ \ the \ sample \ \ j \ in \ \ the \ data \ matrix \ P \ corresponds \ to \ the \ \ author \ \ \ i$$
$$0 \ \ otherwise$$

The neural network is created with appropriate parameters and the data matrices. The articles of TVK and SI authorship is also presented in the form of a Test matrix (X) consisting of 24 function words variables, normalized, in the form of column vectors, each column corresponding to one sample article of TVK and SI authorship. The entire analysis was also performed on the 24 normalized function words variables.

The Neural network (Figure 1.0) was formed and trained using neural network training tool (Graphical User Interface) in MATLAB. The Data were separated randomly into training, validation and testing using divider and command. The analysis of the results obtained indicated that maximum accuracy of classification was achieved 100 percentage of three authors using scaled conjugate gradient algorithm MB, TVK and SI (Table 1.0).

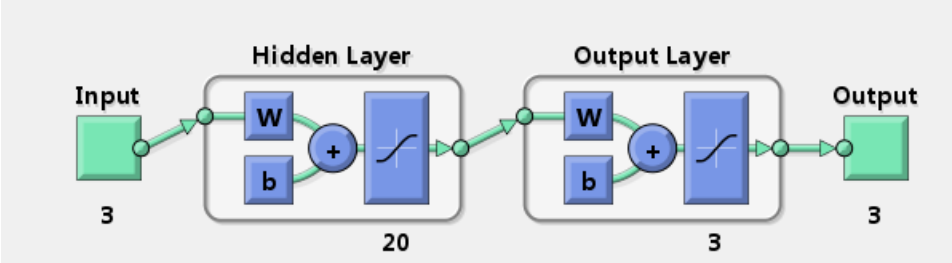

Figure 1.0 *Neural Network for Authorship Attribution*

Table 1.0 *Percentage of Accuracy of MB, TVK and SI using Scaled Conjugate Gradient algorithm*

Name of the Authors	Sample Size	Accuracy of Classification
Mahakavi Bharathiar (MB)	35	100.0%
T.V. Kalyanasundaranar (TVK)	31	100.0%
Subramania Iyer (SI)	38	100%

Table 2.0 *Trained Neural Network*

Three kinds of Samples	Mean Squared Error (MSE)	Percentage of Error	Training Cycle
Training	1.08228e-7	0	1
Validation	1.05640e-7	0	1
Testing	1.05414e-7	0	1

The above results showed that the three authors MB, TVK and SI for training, validation and testing mean squared errors are all zero, which indicates that these three authors writing styles are different

for MB, TVK and SI in neural network classification. Manimannan G and R. Lakshmi Priya (2014) statistically proved these three authors differ from one another using Multivariate Discriminant Analysis in their studies.

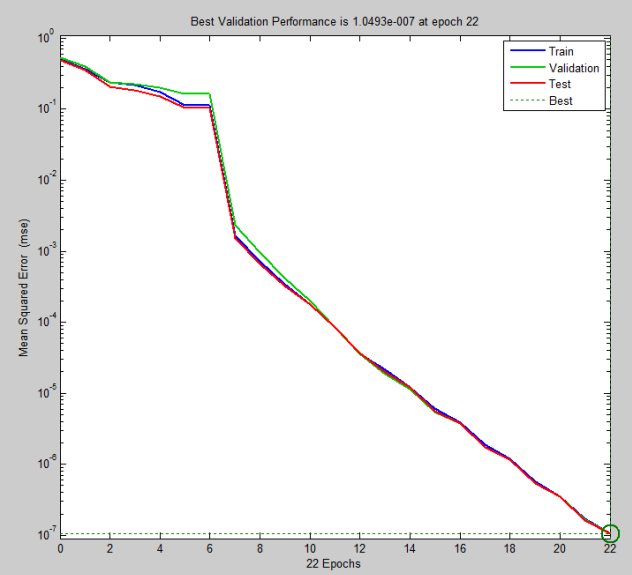

Figure 2.0. *Mean Squared Error (MSE)*

The above results (figure 2.0) showed that the network's presentation improved during training, either click the *Performance* menu in the training tool, or use the function of PLOTPERFORM. Performance is measured in terms of mean squared error and shown in log scale. It quickly reduced as the network was trained. This performance is shown for each of the training, test and validation. The validation data set is little deviated due to testing and training. The report of the network that performed excellently well on the validation set was after training.

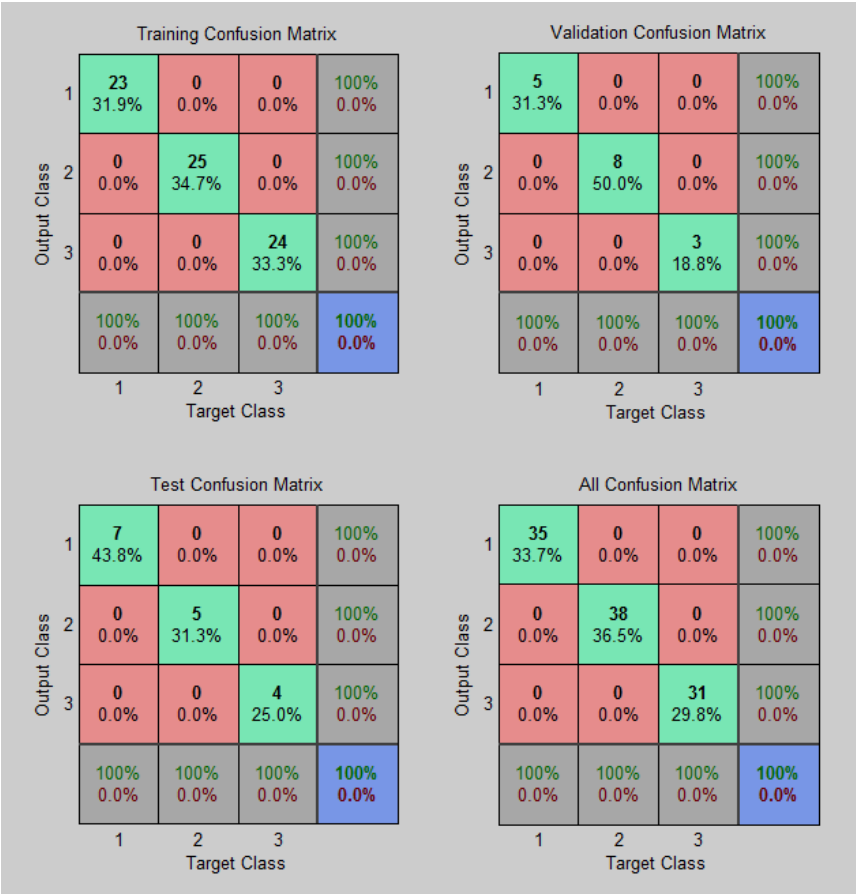

Figure 3.0. *Confusion Matrixes for Training, Validation, Testing and Overall Classification*

The confusion plots of results obtained are shown in figure 3. They indicates that hundred percentage of correctly classified of three authors of MB, TVK and SI data based on the training algorithm used for developing the network. The receiver operating characteristics for three author's classification is shown in figure 4.

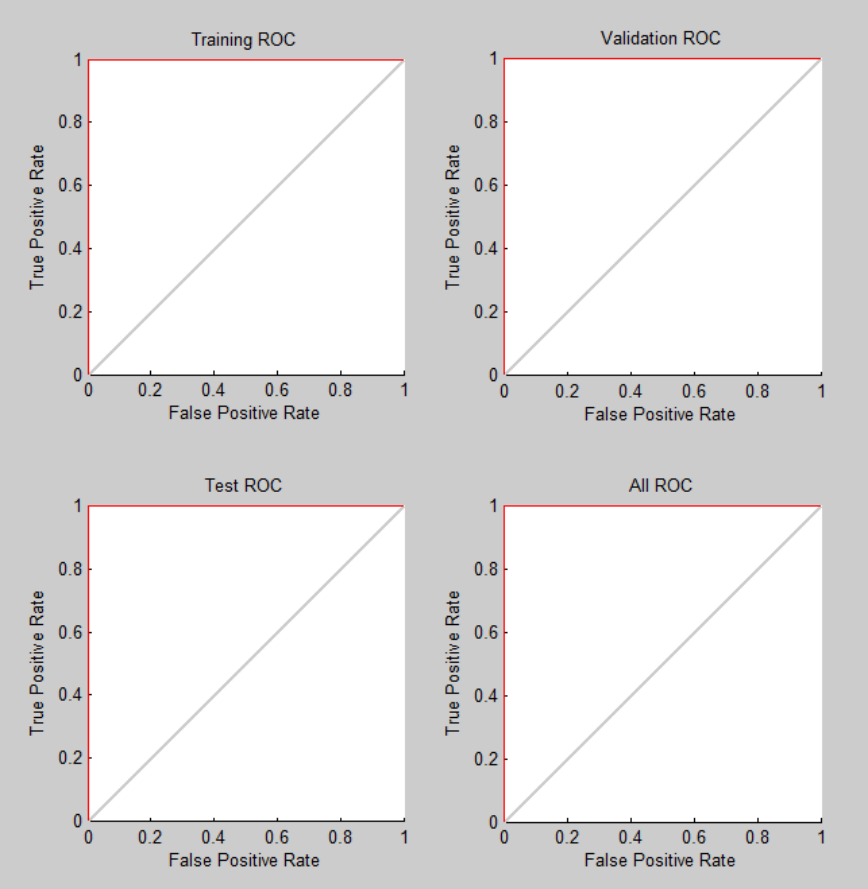

Figure 4.0. *The receiver operating characteristics Curve*

6. CONCLUSION

Application of Artificial Neural Network models has increased considerably in areas of pattern recognition and classification problems in the field of *Authorship* over the last decade. The authorship attribution problem is attempted using a Neural Network Pattern Recognition Tool to use both the supervised and unsupervised learning processes to determine consistent and distinct writing styles of three contemporary scholars. One hundred and four blocks written on India's freedom movement by three scholars in the magazine called *India* in the year 1906 are considered for this study. Considering twenty four function words linguistics variables as stylistic units we have been successful to quantify and classify the witting styles of the scholars to overcome the problems of authorship attribution. All the Blocks of known authorship were hundred percent correctly classified with the articles written by contemporary Tamil scholars, namely Mahakavi Bharathiar (MB), Subramaniya Iyer (SI), and T. V. Kalyanasundaranar (TVK), all of them belonging to the same period.

REFERENCES

[1]. Bagavandas, M. and Manimannan, G. and (2001), Authorship Attribution : The case of Bharathiar, National Conference on Mathematical and Applied Statistics, Department of Statistics, Nagpur University, Nagpur.

[2]. Bailey, R. W. (1979), The Future of Computational Stylistics, Association for Literary and Linguistic Computing Bulletin, Vol. 7, 4-11. England.

[3]. Bishop, C. M (2003), Neural Networks for Pattern Recognition (First Indian Edition), Oxford University Press, New Delhi.

[4]. Chandrasekaran, R. and Manimannan, G. (2008), Neural Network Classification and Authorship Attribution of Articles of Unknown Authorship Using Radial Basis Function, Proceedings of the

National Conference on Artificial Intelligence and Neural Networks, Department of Computer Applications, SRM University, Kattankulathur, Tamil Nadu, pp.246-256.

[5]. Gose, E., Johnsonbaugh, R, and Jost, S (1997), Pattern Recognition and Image Analysis, Prentice Hall Inc., New Jersey.

[6]. Heykin, S. (2001), Neural Networks: A Comprehensive Foundation (Second Edition), Pearson Education (Singapore), New Delhi.

[7]. Holmes, D. I. (1995), The Analysis of Literary Style: A Review, Journal of Royal Statistical Society, Series A, Vol. 148, 328-334, England.

[8]. Holmes D.I. and Forsyth, R.S. (1995), The Federalist Revisited: New Directions in Authorship Attribution, Literary and Linguistic Computing, 10, 111-127, England.

[9]. Kjell, B. (1994), Authorship Determination Using Letter-pair Frequency Features with Neural Network Classifiers, Literary and Linguistic Computing, Vol.9, 119-124, England.

[10]. Authorship Attribution using Tamil Morphological Parameters with Neural Network Classifiers, *International Journal of Scientific and Innovative Mathematical Research (IJSIMR) Volume 2, Issue 11, PP 844-850 ISSN 2347-307X (Print) & ISSN 2347-3142*

[11]. Martin Riedmiller and heinrich Braun. "A direct Adaptive Method for Faster Back Propagation Learning: The RPROP Algorithm" In-stitute fur logic, Komplexitat und Deduktionssyteme, University of Karlsruhe, W-7500 Karlsruhe, FRG, 1993

[12]. Merriam, T. and Mathews, R. (1993), Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher, Literary and Linguistic Computing, Vol.8, 203-209, England.

[13]. Merriam, T. and Mathews, R. (1994), Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe, Literary and Linguistic Computing, Vol.9, 1-6, England.

[14]. Mosteller, F. and Wallace, D. L. (1968), Inference and Disputed Authorship: The Federalist Papers, Addision-Wesley, Messachusetts.

[15]. Powell, M. J. D. (1992), The Theory of Radial Basis Functions Approximation, in Advances of Numerical Analysis , pp. 105–210 , Clarendon Press, Oxford.

[16]. Ripley, B.D. (1996), Pattern Recognition and Neural Networks, Cambridge University Press.

[17]. Tweedie, F. J, Singh, S., and Holmes, D. I. (1996), Neural Network Applications in Stylometry. The Federalist Papers, Computers and the Humanities, 39(1), 1-10, 1996.

[18]. Wasserman, P.D (1989), Neural Computing: Theory and Practice, Von Nostrand Reinhold, New York.

[19]. Wasserman, P.D (1993), Advanced Methods in Neural Computing, Von Nostrand Reinhold, New York.

AUTHORS' BIOGRAPHY

R. Lakshmi Priya, received her M. Sc. M. Phil. in Statistics from University of Madras, Chennai, India. She is Working as Assistant Professor in Statistics, Department of Statistics, Dr. Ambedkar Govt. Arts College, Vyasarpadi, Chennai. She has good knowledge in programming languages like, FORTRAN, PASCAL, COBOL, C, C++, VB and SPSS.

G. Manimannan, received his M. Sc. M. Phil. Ph. D in Statistics from University of Madras, Chennai, India. He received PGDCA (Post Graduate Diploma in Computer Application) from Pondicherry University, Pondicherry, India. He has good research experience by working for many Project Guidance and consultation work in application of Statistics. He has published more than thirty one research papers in various national and International journals. He is good in many programming languages like, FoxPro, HTML, COBOL, C, C++, VB, DBMS, SPSS, SYSSTAT, STATISTICA, MINITAB, MATLAB and working knowledge in SAS and R.