International Journal of Scientific and Innovative Mathematical Research (IJSIMR)
Volume 2, Issue 11, November 2014, PP 855-861
ISSN 2347-307X (Print) & ISSN 2347-3142 (Online)
www.arcjournals.org

Authorship Attribution using Tamil Morphological Parameters with Neural Network Classifiers

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

G. Manimannan
Department of Statistics
Madras Christian College
Tambaram, Chennai, India

Abstract: *Neural Networks have recently been a matter of extensive research and recognition. Neural Network application has increased considerable in areas in which we are accessible with a large amount of data and to classify an underlying pattern. Several neural networks are being drawn forth for authorship determination. Learning Vector Quantization (LVQ) is a neural network technique that develops a "codebook" of quantization vectors and makes use of these vectors to encode any input vector. This Paper proposes, to use LVQ, to attribute the un-attributed articles and verify with the results obtained by traditional statistical analysis. In this paper, comparisons are made in the light of mean of eighteen linguistic features from thirty - two articles written in Tamil by three contemporary Tamil scholars of great repute to determine the authorship of twenty-three un-attributed articles pertaining to the same period. This study throws light on the fact that LVQ is a powerful technique for computational stylistics. This paper makes an attempt to identify the distinct stylistic features of three Tamil Scholars Mahakavi Bharathiar (MB), Subrmani Iyer (SI) and T. V. Kalyasundaranar (TVK) of the same period and also tries to quantify the authorship of unattributed article using eighteen morphological linguistic features.*

Keywords: *Authorship Attribution, morphology, Neural Network, Learning Vector Quantization, Fisher Linear Disriminant Function.*

1. INTRODUCTION

In the past several decades, a wide variety of approaches had been proposed to achieve the perfect classification of un-attributed articles by driving computers into the field. Recently, the emerging technology of neural network has largely exploited to implement a system towards *classification* and *clustering*. In present days, several benchmark and comparison studies had been published on neural network and statistical classifiers. One of the most extensive was the Statlog project (Michie, Spiegelhalter and Taylor, 1994) in which statistical methods, machine learning, and neural networks were compared using a large number of different data sets. As a general conclusion of that study, good neural network classifiers included Learning Vector Quantization (LVQ) technique.

In this research paper, we present the first version of learning vector quantization, a supervised learning neural network model to classify the un-attributed articles and we have shown how potent it is in the discriminative power and the capability to learn and represent implicit knowledge of the computational stylistic data set. Recently, scores of researchers have dug deep to investigate the links between neural network based techniques and traditional statistical classification techniques. In vast majority of cases, these two techniques have been seen as alternatives, or in fact, neural networks, in general, have been seen as a subset of statistics. This approach has led, on one hand, to a fruitful analysis of existing neural networks, and on the other, it has enriched the current statistical methods with newly traced viewpoints, and sometimes has pioneered to a useful synthesis of the two fields.

While working out this Paper, supervised classification using neural network method has been taken into account for authorship determination of the un-attributed articles. In statistical literature, supervised classification is often called discriminant analysis. In this analysis the performance of neural network methods has been discussed in relation to the classification of the

un-attributed articles and this model has been found matching optimally in the design of authorship determination.

In recent years, many scholars have successfully demonstrated that this technique of machine learning field can be applied to authorship attribution. Merriam and Mathews (1994) have trained a multi layer perception network to distinguish the works of Shakespeare and Marlowe. Tweedie *et al.* (1996) have used this machine learning package for reanalysis of the Federalist *Papers*. Kjell (1994) have taken up authorship study using letter-pair frequency features with neural network classification.

2. DATABASE

Statistical stylistic study not only compliments the traditional scholarship of literary experts but also provides an alternative method for investigating the works of doubtful provenance (Holmes, 1998). These Studies provide authentic results if they are worked out within the same genre and also within as close a time period as possible. These stylistic studies inhabit two types of problems, the first being the selection of suitable set of stylistic variables and the second being the selection of appropriate techniques. There is no general agreement on the stylistic variables used in stylistic studies. In general, when choosing the stylistic variables, one must use something that has large variation across authors and relatively little variation among an author's own works.

The present research paper deals with the attributed articles of contemporary Tamil Scholars, namely Mahakavi Bharathiar (MB), Subramaniya Iyer (SI), and T. V. Kalyanasundaranar (TVK), and some un-attributed articles. During the Pre-Independence days, these three Scholars had written number of articles on India's Freedom Movement for publication in the magazine called, India. Initially, all the three contributed their articles by attributing their names. The oppressive attitude of the then British Regime compelled all the three writers to pen articles on the same theme for anonymous publications in the same Magazine. All the attributed and un-attributed articles written on India's Freedom Movement and published in that Magazine were compiled and brought out in 1975 as a book entitled *Bharathi Dharisanam.*

During the study, for authorship attribution, nineteen articles of Bharathi, seven of Subramaniya Iyer, six of Kalyanasundaram, and twenty-three un-attributed articles on India's freedom movement published in 1906 have been technically treated. Each sentence has been categorized as a valid sample. To normalize the stylistic feature, the raw frequency counts of each stylistic feature has been divided by the number of words in each sentence and then multiplied by hundred to express it in percentage. Eighteen stylistic features have been identified in each sentence. The exact lists of variables of this study are given in *Table 1*.

Table 1. *List of eighteen Stylistic Features*

S. No.	Variable Names	S.No.	Variable Names
1	பெயர்ச்சொல் - Noun	10	வினைச்சொற்கள்-Verbs
2	அதிகப்படுத்தி - Intensifier	11	அசைகள்-Syllables
3	குறைவுபடுத்து-Infinitive	12	பின்னொட்டு சொற்கள் -Post-positions
4	பிரதிபெயர்ச் சொல் -Pronoun	13	இடைச்சொல் -Clitics
5	எண்கள்-Numerals	14	வேற்றுமை உருபுகள்-Case Markers
6	இரண்டு எழுத்து சொல் -Two-letter words	15	வினையுரிச்சொல் -Adverb
7	மூன்று எழுத்து சொல்-Three-letter words	16	சேர்க்கும் சொல்-Conjunction
8	நான்கு எழுத்து சொல் - Four-letter words	17	காலங்கள்-Tense
9	உயிரெழுத்துக்களில் துவங்கும் சொல் -Word starting with vowels	18	வார்த்தைகளின் ஒலி -Voice

A Cluster analysis of the nineteen articles 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, 2010). Thus, each article is converted as a raw data matrix and these raw data matrices form the

basis for this data description. Hence the nineteen articles of Mahakavi Bharathiar (MB) consist of three hundred and fifty-three sentences, the seven articles of Subramaniya Iyer (SI) consist of three hundred and fifteen sentences and six articles of T. V. Kalyanasundaranar (TVK) consist of three hundred and eighty-two sentences. The data matrix of MB is of size (353*18), of SI is of size (315*18) and of TVK is of size (382*18) in the case of morphological variables.

3. NEURAL NETWORK APPROACH

One trait that stylometry shares with other applications are that it is fundamentally a case of pattern recognition. However, in most cases of disputed authorship we do not know what the pattern is and we probably would be able to recognize it even if we did. Neural networks have the ability to recognize the underlying organization of data, which is of vital importance for any pattern reorganization problem. Tweedie e*t. al.* (1996) proposed to their application in stylometry will be useful 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.

3.1. Self-Organizing Neural Networks

There are many different types of self-organizing neural networks; however, they all share a common characteristic. This has the ability to assess the input patterns presented to the network, organize itself to learn, on its own, similarities among the collective set of inputs, and categorize (or cluster) them into groups of similar patterns. A special class of self-organizing neural network is based on *competitive learning*. In competitive networks, the output neurons compete among themselves to determine a winner. There are three basic types of self-organizing neural networks. They are:

(i) Kohonen Self-Organizing Map (SOM)
(ii) *Learning Vector Quantization (LVQ) and*
(iii) Adaptive Resonance Theory (ART) networks.

The basic idea of this type of self-organizing neural networks is that the inputs (from a primary event space) are received by a simple network of adaptive elements. The signal representations are mapped (automatically) onto a set of outputs in such a manner that the response attains the same topological order as that of the primary events. Therefore, the network can achieve an automatic formation of topological correct maps of features of observable events.

3.2. Learning Vector Quantization (LVQ)

Learning vector quantization was developed by Kohonen (Kohonen, 1986), and in the year 1990, he summarizes three versions of the algorithm (Kohonen, 1990). This is a supervised learning technique that can classify input vectors based on vector quantization. The version of LVQ presented here is LVQ1, which was Kohonen's first version of learning vector quantization (Kohonen, 1986). The LVQ1 training process proceeds with an input vector being randomly selected (along with the correct class for that vector, thus the supervised learning) from the *labeled* training set. Given an input vector u_i to the network, the "output neuron" (i.e., the class or category) in LVQ1 is deemed to be a *winner* according to

$$\min_{\forall_j} d(u_i - w_j) = \min_{\forall_j} d\left|u_i - w_j\right|^2 \tag{1}$$

3.3. Algorithm

Let u_i for $i = 1, 2, ..., n$ be the set of input vectors, and the network synaptic weight vectors (Voronoi vector) are denoted as w_j for $j = 1, 2, ..., m$. We also let Cw_j be the *class* that is associated with the weight vector w_j and Cu_i is the *class label* of the input vector u_i to the network. The weight vector w_j is adjusted in the following manner:

(i). If the class associated with the weight vector and the class label of the input are the same, that is, $Cw_j = Cu_i$, then

$$w_j(k+1) = w_j(k) + \mu(k)\left[u_i - w_j(k)\right] \tag{2}$$

Where $0 < \mu(k) < 1$ (the learning rate parameter)

(ii). But if $Cw_j \neq Cu_i$, then $w_j(k+1) = w_j(k) - \mu(k)\left[u_i - w_j(k)\right]$ (3)

and the other weight vectors are not adapted.

Therefore, the update rule for modifying a weight vector in 2 is the standard one if the class is correct. In other words, according to the learning rule in 2, the weight vector w_j is moved in the direction of the input u_i if the class labels of the input vector and the weight vector agree. However, if the class is not correct, the weight vector is moved in the opposite direction away from the input vector according to 3. The learning rate parameter $\mu(k)$ is monotonically decreased in accordance with the discrete-time index k (e.g., linearly decreased in time, starting at 0.01 or 0.02; However, many times 0.1 is used as the initial value).

The convergence properties of LVQ have been studied by Baras and LaVigna (Baras and LaVigna, 1990), and their approach is based on stochastic approximation theory. The weights can be initialized by using several methods. In this Paper the first m (total number of classes) vectors from the set of training vectors are used to initialize the weight vectors, that is, $w_j(0)$ *for* $j = 1, 2, ..., m$. Another approach is to randomly initialize the weight vectors (within the dynamic range of the input vectors).

The stopping condition can be based on the total number of desired training epochs, or on monitoring the convergence of the weight vectors. Another stopping condition can be based on monitoring the learning rate parameter directly, and when it is sufficiently small, training can be terminated. In this Paper, we establish the stopping condition to be the total number (predefined) of iterations. The basic LVQ algorithm can be summarized as follows:

Step 1. Initialize all weight vectors $w_j(0)$, initialize the learning rate parameter $\mu(0)$), and set $k = 0$.

Step 2. Check the stopping condition. If false, continue; if true, quit.

Step 3. For each training vectors u_i, perform step 4 and 5:

Step 4. Determine the weight vector index $(j = q)$ such that

$\min\left|u_i - w_j(k)\right|^2$ and the weight vector $w_q(k)$ that minimizes the square of the norm.

Step 5. Update the appropriate weight vector $w_q(k)$ as follows:

$$If \quad \mathrm{Cw_q} = Cu_i, then \;\; w_q(k+1) = w_q(k) + \mu(k)\left[u_i - w_q(k)\right]$$

$$If \quad \mathrm{Cw_q} \neq Cu_i, then \;\; w_q(k+1) = w_q(k) + \mu(k)\left[u_i - w_q(k)\right]$$

Step 6. Set $k \leftarrow k+1$, reduce the learning rate parameter, then go to **Step 2.**

The learning rate parameter μ can be reduced in accordance with k (discrete–time index)

sing $\mu(k) = \mu(k+1)(k+1)$ *for* $k > 0$ μ(k) = μ (k + 1)/(k + 1) for k > 0.

The neural network architecture for LVQ is shows in *Figure 1*

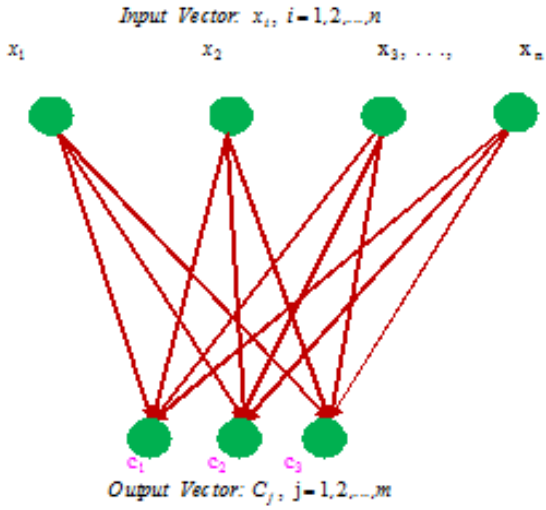

Figure 1. *Neural Network architecture for Learning Vector Quantization*

4. RESULT AND DISCUSSION

In the present research paper discussing nineteen articles of MB to consider as class one, seven of SI to class two and six of TVK to class three, and thereby we have totally thirty two vectors associated with three classes. The following algorithm developed for LVQ analysis.

Step 1: To initialize the codebook vectors w_1, w_2, and w_3, is to use the first three articles, one from each group of articles. The associated classes for w_1, w_2, and w_3 are coded 1, 2, and 3 respectively.

Step 2: The remaining twenty-nine articles vector matrix of the three Tamil scholars can be used for training.

Step 3: To initialize the learning rate parameter to $\mu(k=1) = \mu(1) = 0.1$ and decrease it by k every training epoch, for example $\mu(2) = \frac{\mu(1)}{2}, \mu(3) = \frac{\mu(2)}{3}$, etc.

Step 4: According to the LVQ algorithm, we set $k = 2$, and $\mu(2) = \frac{\mu(1)}{2} = 0.05$, then go to Original LVQ algorithm step 2 and check the stopping condition.

Step 5: The next training epoch we start with the first training article among the remaining twenty-nine articles of the three scholars. The MATLAB program is used with the number of training epochs set to 20,000 to find the Codebook vectors.

Step 6: After 20,000 training epochs, the final weights give the following *codebook* vectors in *Table 1.*

Table 1. *Codebook vectors after 20,000 training epochs.*

w_1	w_2	w_3	w_1	w_2	w_3
34.7039	45.0577	41.4376	23.5462	21.5966	21.9843
0.3400	5.8963	5.5785	257.6893	233.5009	245.8277
0.6358	1.8456	3.8853	13.0731	38.2357	30.3212
8.1532	9.2301	6.6337	14.1369	34.2626	32.0344
3.6737	5.3992	5.6836	38.8306	76.3359	61.3238
10.4447	9.5503	9.8001	4.0836	2.4753	2.6011
19.0042	18.7133	18.2916	22.5499	45.1894	36.7108
20.0785	25.9585	25.0743	1.6926	4.8204	1.8052

27.4689	32.9134	29.6252	2.3332	2.0338	1.6514

The convergence profile for the elements of the training epoch is shown in figure 2-4. Finally, if we calculate the minimum distance between each of the unattributed articles and the computed weight vectors w_1, w_2, and w_3 of the codebook shown above, it will give the class to which each unattributed articles belongs, and it is computed using the MATLAB code. The experiment reveals that all the twenty-three unattributed articles have been written by the Tamil Scholar Mahakavi Bharathiar.

Figure 5 shows the performance of the LVQ Network after training. The final classification results based on Fisher Linear Disriminant Function matrix and LVQ codebook vectors of the attributed and un-attributed articles have been shown in figure 6 and 7.

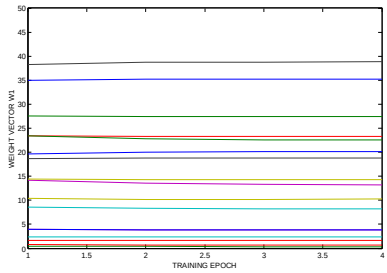

Figure 2. *Convergence profiles for the elements of weight vector* w_1

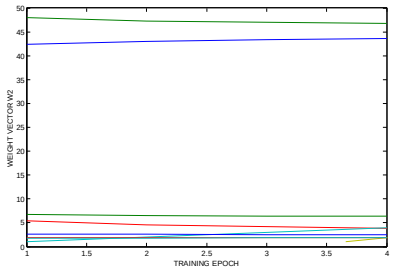

Figure 3. *Convergence profiles for the elements of weight vector* w_2

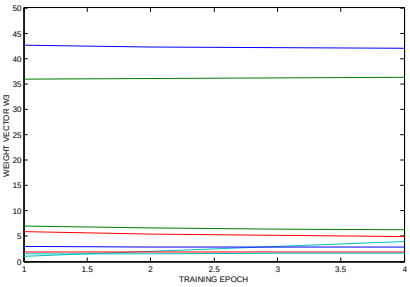

Figure 4. *Convergence profiles for the elements of weight vector* w_3

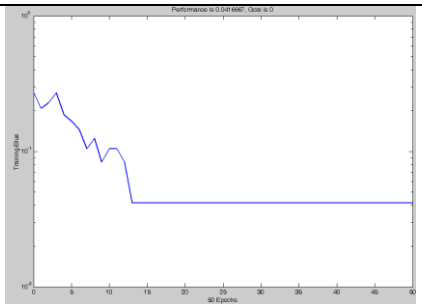

Figure 5. *LVQ Network after training*

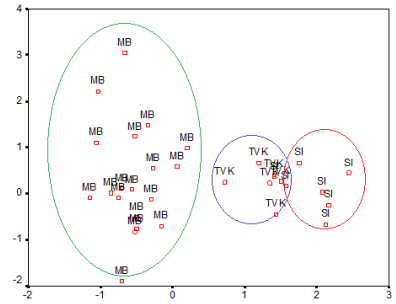

Figure 6. *Classification Map of MB, SI and TVK using Fisher Linear Disriminant Matrix*

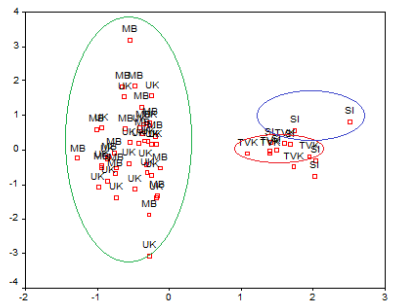

Figure7. *Classification Map of MB, SI, TVK and Unattributed Articles using Fisher Linear Disriminant Matrix*

5. CONCLUSION

Neural network method for classification and clustering of computational stylistic data sets has been discussed both from a theoretical point of view and in the context of experiments. Neural networks are online learning systems, intrinsically nonparametric and model-free, with the learning algorithms typically of the error correction type. Further, the power of *neural* algorithms amplified the flexibility of architecture and the possibility for incremental learning. It is simple and local, and avoids heavy numerical operations that make use of a fixed set in batch mode. As an experiment, we have considered the computational stylistic data sets of attributed and un-attributed articles and examined all the twenty-three unattributed articles. The results of this study

bring to lime light the fact that all the twenty-three unattributed articles had been written by the Tamil scholars, Mahakavi Bharathiar.

REFERENCES

[1] Michie, D., Spiegelhalter, D.T., and Taylor, C.C., Eds., (1994). *Machine Learning, Neural, and Statistical Classifications.* London: Ellis Horwood Ltd.

[2] Nasrabadi, N. M., and King, R. A., (1988). "Image Coding Using Vector Quantization: A Review", *IEEE Transactions on Communications,* vol. 3, pp. 957-71.

[3] Merriam, T. and Mathewa, R. (1994). Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Mralowe. *Literary and Linguistic Computing,* 9:1-6.

[4] Tweedie, F. J, Singh, S., and Holmes, D. I. (1996a). Neural Network Applications in Stylometry. The Federalist Papers. *Computers and the Humanities*, 30: 1- 10.

[5] Kjell, B. (1994). Authorship Determination Using Letter-pair Frequency Features with Neural Network Classifiers. *Literary and Linguistic Computing*, 9: 119-24.

[6] Holmos, D. I., (1998). The Evolution of Stylometry in Humanities Scholarship, *Literary and Linguistic Computing,* vol. 13, pp. 11-117.

[7] Manimannan, G. and Bagavandas, M. (2010). Identification of Consistent and Distinct Writing Styles: A Clustering-based Stylistic Analysis, *Language Forum, A Journal of Language and Literature*, New Delhi, pp. 57-72.

[8] Kohonen, T., (1986). "Learning Vector Quantization for Pattern Recognition", *Technical Report TKK-F-A601*, Helsinki University of Technology, Finland.

[9] Kohonen, T., (1990). "Improved Version of Learning Vector Quantization", *Proceedings of the International Joint Conference on Neural Networks*, San Diego, CA, vol. 1, pp. 545-50.

[10] Kohonen, T., (1990). "The Self-Organizing Map", *Proceedings of the Institute of Electrical and Electronics Engineers,* vol. 78, pp. 1464-1480.

[11] Baras, J. S., and LaVigna, A., (1990). "Convergence of Kohonen's Learning Vector Quantization", *International Joint Conference on Neural Networks,* San Diego, CA, vol. 3, pp. 17-20.

AUTHORS' BIOGRAPHY

R. Lakshmi Priya received his 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++, 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 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.