PAPER *Special Issue on Nonlinear Theory and its Applications*

A Neuro Fuzzy Algorithm for Feature Subset Selection

Basabi Chakraborty[†] *and* **Goutam Chakraborty**[†], *Nonmembers*

SUMMARY

Feature subset selection basically depends on the design of a criterion function to measure the effectiveness of a particular feature or a feature subset and the selection of a search strategy to find out the best feature subset. Lots of techniques have been developed so far which are mainly categorized into classifier independent *filter* approaches and classifier dependant *wrapper* approaches. Wrapper approaches produce good results but are computationally unattractive specially when nonlinear neural classifiers with complex learning algorithms are used.

The present work proposes a hybrid two step approach for finding out the best feature subset from a large feature set in which a fuzzy set theoretic measure for assessing the goodness of a feature is used in conjunction with a multilayer perceptron (MLP) or fractal neural network (FNN) classifier to take advantage of both the approaches. Though the process does not guarantee absolute optimality, the selected feature subset produces near optimal results for practical purposes. The process is less time consuming and computationally light compared to any neural network classifier based sequential feature subset selection technique. The proposed algorithm has been simulated with two different data sets to justify its effectiveness.

key words: feature subset selection, neuro fuzzy approach, fuzzy measure, feature ranking, fractal neural network

1. Introduction

Selection of a good subset of features from a large set of features is an important preprocessing task of an automatic pattern recognition system. The main objective of feature selection is to retain the optimum discriminatory characteristics necessary for the recognition process and to reduce the dimensionality of the measurement space so that effective and easily computable algorithms for classification can be devised. To facilitate the selection process, the quality of any feature has to be assessed via some well designed criterion function. More important task is to find out the best feature subset from a large number of subsets as it is well known that the best two individual features may not comprise the best feature subset of two features.

General feature subset selection techniques [1] are based on the design of a *criterion function* and the selection of a *search strategy*. The criterion function determines the suitability of one feature subset over another while the search strategy decides the best possible feature subset among a number of candidates. Researchers have developed a lot of criterion functions and search strategies, mostly from the statistical theory and they have their relative merits and demerits. The existing approaches of feature subset selection can also be viewed as broadly classified into two categories: *filter* and *wrapper* approaches. Filter approaches [2] [3] base on a criterion function which is classifier independent while wrapper approaches [4] use the classifier accuracy as the criterion function and depends on the learning algorithm of the specific classifier. The two approaches have basic merits and demerits. While classifier dependant methods produce good results, specially when the classifier is designed to solve the particular problem, it is not computationally attractive when the number of input features are large. The computational burden is more heavy when nonlinear neural classifier with complex learning algorithms are used.

In this work a hybrid two stage approach to the problem of feature subset selection has been undertaken to take advantage of both the approaches. In the first stage, a fuzzy set theoretic criterion function developed by the author [5] previously, has been used to assess the quality of a particular feature from a set of features independent of any classifier. The features are then ranked according to their effectiveness measured by the criterion function and some of the tail ranked features are removed from the set of features. In the second stage, an artificial neural network classifier has been used with the reduced set of ranked features to determine the best feature subset. Two types of connection structure, full connection and statistically fractal connection, have been used for the neural classifier. The algorithm has been simulated by two different data sets. Though the method may not find the best feature subset in the strict sense of optimality, the algorithm pruduces near optimal feature subset in a re asonable time lesser than any wrapper approaches with neural classifier [6].

The neural classifiers used in this work have been described in the next section. The proposed feature subset selection algorithm in details is presented in the section 3. The Section 4 represents the simulation of the proposed algorithm by the popular iris data set used for pattern classification problems and by the sonar data set in target recognition problem. Section 5, the final section, contains conclusion.

[†]The authors are with Faculty of Software and Information Science, Iwate Prefectural University, 152-52 Azasugo, Takizawamura, Iwate 020-0193, Japan

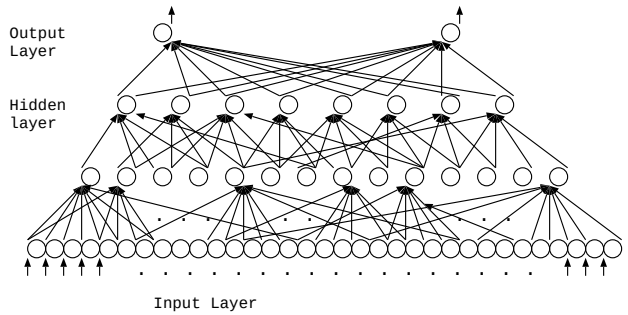

Fig. 1 The architecture of the proposed fractal net.

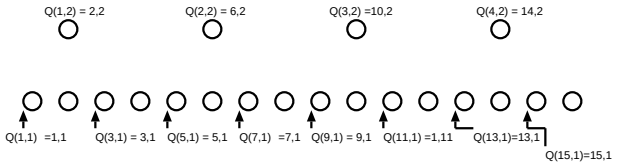

Fig. 2 Spatial Positions of the Upper and Lower layer Neurons

2. Neural Network Classifier

Artificial neural networks are now very much popular for using in pattern classification problems. The most common architecture of artificial neural network classifier is multilayer perceptron(MLP) with back propagation learning algorithm. In our work popular MLP classifier and a fractal neural network classifier (FNN) proposed earlier by author in [7] for pattern classification problems, have been used for the selection of feature subset. Fractal neural network classifier is a modified version of MLP in which the connection structure of neurons within layers follows a power law that generates a statistically fractal and sparse connection structure. The architecture is described in the following subsection.

2.1 Fractal Neural Network

The fractal neural network model is a modified version of feedforward multilayer neural network in which upper layer neurons are connected to the lower layer neurons with a probability following an inverse power law which generates a sparse network with statistically fractal [8] connection structure. However the final hidden layer is fully connected to the output layer. Each layer is an array of neurons in one or two dimension depending on the type of input to be processed. The probability that ith processing element in the kth layer receives connection from the jth processing element of the previous layer, defined by CP_{ijk}

follows the law

$$\begin{aligned} CP_{ijk} &= Ar_{ijk}^{D_k-d} \\ i &= 1,2\ldots n_k \\ j &= 1,2\ldots n_{k-1} \\ 0 &\leq D_k \leq d \end{aligned} \tag{1}$$

where r_{ijk} is the Euclidean distance between ith processing element in the kth layer (considering one dimensional layers) and jth processing element of the previous layer defined as

$$r_{ijk} = ||Q_{ik} - Q_{j(k-1)}||,\ r_{ijk} \geq 1 \tag{2}$$

d denotes dimension of the array of neurons in kth layer. A represents a constant, D_k represents the fractal dimension (similarity dimension) of the synaptic connection distribution of k th layer. Q_{ik}, and $Q_{j(k-1)}$ denotes the spatial position of the ith processing element in the kth layer and jth processing element of the previous layer defined by

$$Q_{ik} = [\lceil n_{k-1}(2i-1)/2n_k \rceil, k] \text{ for } i = 1,2,\ldots,n_k \tag{3}$$

where n_{k-1} and n_k represents the number of neurons in the $(k-1)$th and kth layers respectively. Figure 1 represents the architecture of the proposed model for one dimensional layers. Figure 2 explains the positioning of neurons according to Equation 3 in case of 16 and 4 neurons in the lower and upper layer respectively.

To implement such a sparse NN, for each i, j, k, a uniform random number ρ in the interval [0,1] has to be generated and the connectivity C_{ijk} of the link from the ith processing element in the kth layer to the jth processing element of the previous layer is to be assigned as

$$\begin{aligned} C_{ijk} &= 1, \text{ if } CP_{ijk} \geq \rho \\ &= 0, \text{ Otherwise} \end{aligned} \tag{4}$$

Initially the weights of the connected links are initialized with random values in a range appropriate for the particular problem. The network is trained in a supervised mode by error backpropagation where the weights of the connected links only are gradually adjusted to match the teacher output with actual output.

The connection structure of the network allows low probability of long range connection links and high probability of short range connection links. This sparse architecture has been proved effective in pattern classification problem compared to fully connected perceptron specially when the data set contains redundant information [7].

3. Feature Subset Selection Algorithm

A hybrid two stage neuro-fuzzy feature subset selection scheme for a N feature $(F = F_1, F_2, \ldots F_N)$, C class pattern classification problem has been proposed in this work.

The proposed feature subset selection algorithm follows two steps and is described in the following two subsections.

3.1 Fuzzy set theoretic Measure in Feature ranking

In the first step, effectiveness or quality of all the features are evaluated by a fuzzy set theoretic criterion function and the features are ranked according to their goodness. The details of the criterion function and its use for feature ranking has been reported earlier in [5] and is described here in short. The basic logic behind the design of the *feature evaluation index* for assessing effectiveness of a particular feature in the context of a two-class recognition problem is that the *index* should represent a measure of uncertainty/ambiguity or information associated within the class and between the classes when represented with that particular feature. The feature which represents the classes with minimum intraclass uncertainty/ambiguity and maximum interclass uncertainty/ambiguity is noted as the best feature. Accordingly, *Feature Evaluation Index (FEI)* of a particular feature has been defined as the ratio of intraclass to interclass ambiguity measures. The measure of fuzziness such as *index of fuzziness* and *entropy* are used to measure the intraclass and interclass fuzziness/ambiguity. The particular feature is considered to be efficient or good when the intraclass fuzziness value is low and the interclass fuzziness value is high, making FEI lower for better feature.

FEI for qth feature F_q is defined mathematically as

$$(FEI)_q = \frac{d_{qj} + d_{qk}}{d_{qjk}} \tag{5}$$

$$j, k = 1, 2, \cdots, C \text{ and } j \neq k,$$
$$q = 1, 2, \cdots, N$$

where d_{qj} and d_{qk} represent the fuzzy measure of intraclass ambiguity value of jth and kth class and d_{qjk} represents the fuzzy measure of interclass ambiguity value of jth and kth class pooled together to be considered as a single class when the classes are represented by qth feature only. For a multiclass problem, the average of the $(FEI)_q$ values of all possible pair of classes (j and k) are taken as the measure of goodness of q th feature.

The ambiguity or fuzziness of a fuzzy set is represented by various fuzzy measures like *index of fuzziness* or *fuzzy entropy* [9]. According to Delucca and Termini *Fuzzy entropy* of a fuzzy set A with n supporting points is defined as

$$H(A) = \frac{1}{n \ln 2} \sum_i S_n(\mu_A(x_i)), i = 1, 2, \ldots, n \tag{6}$$

with Shanon's function
$S_n(\mu_A(x_i)) = -\mu_A(x_i) \ln \mu_A(x_i)$

$$-\{1 - \mu_A(x_i)\} \ln\{1 - \mu_A(x_i)\} \tag{7}$$

where $\mu_A(x_i)$ denotes the membership value of the point x_i in the fuzzy set A. Membership values can be assigned by standard π or S function [10].

Here fuzzy entropy has been used as the measure of ambiguity for evaluating FEI given by Eq 5. The features are then ranked according to FEI values, lower value corresponds to higher rank. Depending on the problem and the number of features and their FEI values, a number of top ranking features are retained for further processing in the second step.

3.2 Neural Network Classifier for Feature Subset Selection

In the second step artificial neural network has been used to find out the best feature subset. We have used two neural classifiers, popular MLP and FNN having a fractal connection structure as described in the previous section. We have also used two different procedures, simple near optimal algorithms, to find out the optimum feature subset which are described below.

1. The first algorithm which we have proposed in [11] have been used here with the reduced set of features selected in the first step in order to have an improvement over time from the previous one stage sequential selection of features in [11] to the present two stage process with the initial knowledge of feature ranking. The algorithm is as follows:

   a. The neural network classifier has been set up with the number of input neurons same as the number of features, number of output neurons as the number of classes. Number of hidden layers, number of neurons in the hidden layers and fractal dimension in case of FNN are heuristically selected depending on the optimum performance of the network.
   b. The network has been trained for optimum efficiency determined by the highest classification rate tested by the test samples by suitable choice of the parameters.
   c. The network is initially set for all the features and then features are removed one at a time to examine the effect of its removal on the optimum classification rate. The feature with the least effect has been selected for final removal.
   d. The last step is continued for a number of times to remove a selected feature from the feature set at each time until the optimum number of features is reached determined by a pre assigned stopping criterion depending either on limit of classification error or number of features in the finally selected feature subset.

   The above algorithm has been used with both the classifiers (MLP and FNN) to find out the best feature subset and their performance and results have been compared.

2. The second algorithm is more simpler and less time

consuming but optimal selection is not guaranteed. Here the features, ranked in the previous stage are fed to the input layer according to the order of ranking. In case of FNN, fractal dimension of the connection structure of the nodes in the second layer is varied to take a higher to lower value for different nodes. The actual algorithm is described below.

a. Various parameters of the neural network classifier are set as usual like the step one of the first algorithm.

b. The network is then trained in such a way that produces highest classification score while tested with the test samples.

c. In case of FNN, for the same set of values of fractal dimension, several fractal models with different connection structure are trained similarly for optimum efficiency tested by highest classification score of test samples.

d. The most efficient model for both MLP and FNN, measured by the best classification rate, is to be selected. At this stage, the network connection weights are examined. Connection weights smaller than a pre-assigned value depending on the problem have been considered as zero, that is, no connection.

e. The remaining connections are examined for finding out the best feature subset. The inputs which are connected to a particular output having value of the connection weight greater than a preassigned limit and connected to other outputs having values of the connection weights smaller than a preassigned limit are selected as the discriminatory inputs for that particular output.

As the inputs represent the features and the outputs represent the classes, it can be intuitively inferred that the method selects the discriminatory features for a particular class. Similarly examining all the outputs we can find out the discriminatory inputs for each of them by examination of the connection weights. The superset of all these discriminatory input feature sets will comprise the near optimal best feature subset for the multiclass pattern classification problem.

4. Simulation and Results

The proposed feature subset selection scheme has been tested by simulation with two different data sets. One of them is Anderson's IRIS data set [12], commonly used to test pattern recognition problems and the other one is SONAR data set used for underwater target recognition [13] problem.

Features	Feature no.	Rank no.
F_1	1	13
F_2	2	15
F_3	3	1
F_4	4	2
$F_1 - F_2$	5	12
$F_1 - F_3$	6	7
$F_1 - F_4$	7	16
$F_2 - F_3$	8	3
$F_2 - F_4$	9	4
$F_3 - F_4$	10	11
F_1/F_2	11	14
F_1/F_3	12	6
F_1/F_4	13	10
F_2/F_3	14	9
F_2/F_4	15	5
F_3/F_4	16	8

Table 1 Generated features and their ranking according to FEI for IRIS data

Feature subset selection with first algorithm			
by MLP		by FNN	
selected feature subset	time taken	selected feature subset	time taken
3		4	
4	1.52 hrs	8	0 .58 hrs
8		3	
9		9	
12		12	

Table 2 Feature subset selection with first algorithm for IRIS data

4.1 Simulation with IRIS data

This data set contains three classes i,e. three varieties of IRIS flowers, namely Iris Setosa, Iris Versicolor & Iris Virginica each with 50 sample vectors. Each sample has four feature vectors (Sepal length F_1, Sepal width F_2 , Petal length F_3 & Petal width F_4). As the number of features in this case are small, the feature set has been extended and twelve features have been generated from the primary four features and all togther sixteen features are considered as the feature set for our experiment.

In the first step of the feature subset selection process, the features are ranked from good to bad according to FEI values calculated by using Eqs. 5, 6 and 7. The following order represents the features in decreasing order of goodness measure as is also evident from Table 1

$3, 4, 8, 9, 15, 12, 6, 16, 14, 13, 10, 5, 1, 11, 2, 7$

In the second step the features ($3, 4, 8, 9, 15, 12, 6, 16, 14, 13$) are used for final subset selection. MLP and fractal neural network (FNN) both are used separately for feature subset selection. Table 2 and Table 3 represents the finally selected feature subset and the time taken for finding out the final subset for both the networks and for the both the algorithms respectively. Table 4 represents the discriminatory features for all the classes.

Feature subset selection with second algorithm			
by MLP		by FNN	
selected feature subset	time taken	selected feature subset	time taken
3	.27 hrs	3	.15 hrs
4		4	
8		8	
9		9	
15		12	
16		16	

Table 3 Feature subset selection with second algorithm for IRIS data

Discriminatory features for		
class 1	class 2	class 3
8	8	4
3	9	3
9	4	8
4	15	9
12	16	12
10	12	14

Table 4 Feature subset selection from ANN for IRIS data

It is found that the fractal network takes lesser time to find out the final feature subset than MLP though the final subset comes out to be comprised of same features in both types of networks. It is also found that the total time taken for the two stage algorithm is quite less than the one step sequential algorithm proposed in [11]. From both the algorithms it is seen that the most important feature subset is $(3, 4, 8, 9)$ which conforms well with the earlier results obtained for IRIS data.

4.2 Simulation with SONAR data

This data set, collected from underwater target classification problem using sonar signals, consists of two types of sonar returns, one from a metal cylinder and the other from a cylindrically shaped rock, both of them positioned on a sandy ocean floor. The impinging pulse was a wideband linear FM chirp and the returns were obtained from each target at various aspect angles. A set of 208 returns (111 cylinder and 97 rock returns) were selected on the basis of the strength of the specular return (4.0 -15.0 dB signal-to-noise ratio), making certain that a variety of aspect angles were represented. Each sample signal was preprocessed to produce power spectral envelope and 60 sample points were obtained for each envelope. These samples were normalized to take on values between 0.0 and 1.0 for using as the 60 input features.

In the first step, the features are ordered from good to bad according to FEI values, calculated by using Eqs. 5, 6 and 7 and the ordered features are fed to the input of the neural network in the second step. Here, as the number of features are large only the second algorithm has been tried with MLP and FNN.

Both MLP and FNN have been constructed with

No. of features in subset	Time taken		
	for present		for sequential algorithm with FNN
	algorithm with MLP	algorithm with FNN	
10	2.85 hrs	2.32 hrs	5.43 hrs
5	2.85 hrs	2.32hrs	7.41 hrs

Table 5 Time taken for feature subset selection by present algorithm and se quential algorithm for SONAR data

No. of features in subset	Average recognition rate		
	for present		for sequential algorithm with FNN
	algorithm with MLP	algorithm with FNN	
10	93.6%	93.4%	93.5 %
5	93.9 %	93.9 %	93.8 %

Table 6 Feature subset selection from ANN by present algorithm and sequential algorithm for SONAR data

60 neurons in the input layer and 2 neurons in the output layer. For FNN, the connection structure has been set up according to Eq. 1 and Eq. 4 with the values of A and d taken as 1 as before. The number of neurons in the hidden layer is varied between 4 to 24 and the value of the fractal dimension has also been varied (from 0.8 to 0.95) for different nodes in the hidden layer to introduce a gradual sparsity in the network. The high ranked features are fed to the input of the more denser part of the network. The network has been trained with backpropagation algorithm with initial weights selected from random numbers between 0.3 and -0.3. For a particular connection structure different networks with varying initial weight set up have been trained and tested for classification rate with same test samples. Now the network having highest classification rate for test samples have been selected for examining its connection structure in case of both MLP and FNN. In the final step, the connection structure of the selected network has been examined to find out the best feature subset. Following the method described in the previous section the best feature subset of 10 and 5 features are found out from the examination of the connection weights by taking suitable preassigned limit. The time taken for finding out the best feature subset with the present algorithm and the one stage sequential algorithm in [11] have been shown for comparison in Table 5 for both MLP and FNN. Table 6 represents the classification rate with the selected feature subsets and the feature subsets selected by sequential one stage feature selection algorithm for both MLP and FNN.

5. Conclusion

Feature subset selection is very important prior to classification. Most of the collected real data set contains redundant or irrelevant information. Selection of the most discriminatory information is the key to successs of any pattern classification system. In feature selection process, the assessment of an individual feature as well as the assessment of a group of features is needed

as the best two features do not always make the best feature subset of two features.

While statistical techniques to the problem of feature selection and consequently various search techniques for the best feature subset selection are well known and mathematically strong, they are computationally unattractive specially in case of real world large data set problems. Artificial neural networks are nowadays becoming popular as pattern classifier.

In this work a hybrid two stage feature subset selection scheme has been adopted to lessen time and computational burden. In the first stage a fuzzy set theoretic measure has been used to have a preliminary assessment of the goodness of individual features. The measure, already developed by the author for feature ranking, has been shown to be an efficient measure for feature ranking and computationally less complex than well known statistical measures. In the second stage a fractally connected artificial neural network classifier and popular MLP have been used for final feature subset selection. The use of fractal connection structure results in sparseness for which the training time of the network is considerably lower than a fully connected neural network of same dimension.

The use of aprioi knowledge obtained by ranking the features initially and feeding the ordered features as the input to the neural classifier as presented in the second algorithm reduces considerably the time and steps needed for feature subset selection by neural network classifier. The present work reduces the time required for the best feature subset selection compared to the previous work reported in [11] as the number of steps here is substantially reduced by preliminary choice of features using fuzzy criterion function.

It has been noted from the simulation done in this work that the hybrid scheme with the fuzzy criterion function and fractal neural network classifier using the second algorithm is efficient in selecting optimal feature subset for any pattern recognition problem. The scheme is easy, computationally light and less time consuming. Though the second algorithm does not guarantee any optimality for the selected feature subset, simulation results show that the selected subset is good enough for recognition. The presented algorithm in this work seems to be an efficient practical algorithm for finding near optimal feature subset from a large feature set of real world multidimensional pattern recognition problems.

References

[1] P. Devijver and J. Kittler,*Pattern Recognition: A Statistical Approach*, Prentice Hall, 1982.

[2] H. Almuallim and T. G. Dietterich ," Learning boolean concepts in the presence of many irrelevant features", *Artificial Intelligence*, Vol 69, pp. 279–305, 1994.

[3] K. Kira and L. A. Rendell, " The feature selection problem: Traditional methods and a new algorithm" *Proceedings of Ninth National Conference on Artificial Intelligence*, pp. 129–134, MIT Press, 1992.

[4] G. H. John, R. Kohavi and K. Pfleeger, " Irrelevant feature and the subset selection problem", *Machine Learning: Proceedings of the Eleventh International Conference*, pp. 121–129, 1994.

[5] S. K. Pal and Basabi Chakraborty, "Fuzzy set Theoretic Measure for Automatic Feature Selection", *IEEE Trans. on Sys, Man and Cybernatics*, Vol SMC-16, No.5, pp. 754–760, Sep./Oct. 1986.

[6] R. Setiono and H. Liu, "Neural Network Feature Selector", *IEEE Trans. on NN*, Vol. 8, No. 3, pp. 654–662, May 1997.

[7] Basabi Chakraborty, Yasuji Sawada and Goutam Chakraborty, "Layered Fractal Neural Net: Computational Performance as a Classifier", Knowledge-Based Systems,Vol 10, No. 3, pp. 177–182, October, 1997.

[8] B. B. Mandelbrot, "The fractal Geometry of Nature", Freeman, SanFrancisco, CA, 1982.

[9] A. Deluca and S. Termini,"A definition of a nonprobabilistic entropy in the setting of a Fuzzy Sets Theory", *Information and Control*, Vol 20, pp.301–312, 1972.

[10] L. A. Zadeh, "Fuzzy Sets and their Application to Cognitive and Decision Processes", Academic Press, 1975.

[11] Basabi Chakraborty and Yasuji Sawada, "Fractal Neural Network Feature Selector for Automatic Pattern Recognition System", IEICE Trans. on Fundamentals of Electronics, Communications and Computer Sciences,Vol E82-A, No.9, pp.1845–1850, September, 1999.

[12] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Functions", Plenum Press, NY, 1981.

[13] P. R. Gorman and T. J. Sejnowski, "Analysis of Hidden Units in a Layered Network Trained to Classify Sonar Targets", Neural Networks, Vol 1, pp.75–89, 1988.

Basabi Chakraborty received M. Tech and Ph.D degrees in Radio Physics and Electronics from Calcutta University, India. She worked as a part time researcher for two years in AIC Systems Laboratory in Sendai, Japan until 1993. She received another Ph.D in Information Science from Tohoku university, Japan in 1996. Currently she is with the department of Software and Information Science of Iwate Prefectural University, Japan. Her main research interests are in the area of Pattern Recognition, Fuzzy logic, Genetic Algorithm, Artificial Neural Network and Computer Communication Networking.

Goutam Chkraborty received his Ph.D. in Information Engineering in 1993 from Tohoku University, Japan. Presently he is an Associate Professor in the Iwate Prefectural University, Japan in the department of Software and Information science. His research interests are Neural Network, Fuzzy logic, Genetic algorithm etc. and their applications to solve different classification, scheduling and optimization problems including applications to Computer Networks and mobile networking.