Data Classification by using Homogeneous Clustering Process

Bhavana R. Mundane
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India

Swati V. Shinde
Department of Information Technology
Pimpri Chinchwad College of Engineering
Pune, India

ABSTRACT
In this paper we have performed dynamic clustering based on classification. The Enhance Neuro-fuzzy system for classification using dynamic clustering presented in this paper is an extension of the original Neuro-fuzzy method for linguistic feature selection and rule-based classification. In the dynamic clustering process, the parameter values like centroid, threshold value and standard deviation are estimated. This parameter is used for creating the cluster. The Gaussian membership function is applied to these clusters to generate the binary value of each feature to given cluster. Using this method we have got the large number of cluster and minimum accuracy. To reduce the cluster size and to improve the accuracy we have implement the homogenous clustering process. Using this process we have minimize the cluster size, and also improve the accuracy.

General Terms
Fuzzy logic, Artificial intelligence, Neural Network.

Keywords
Neuro-fuzzy, Classification, Dynamic Clustering, Class based grouping, homogenous clustering.

1. INTRODUCTION
Fuzzy logic is a form of many-valued logic that deals with approximate, rather than fixed and exact reasoning. Compared to traditional binary logic, fuzzy logic variables may have a truth value that ranges in degree between 0 and 1 [1]. Fuzzy logic has been extended to handle the concept of partial truth, where the truth value may range between completely true and completely false. Furthermore, when linguistic variables are used, these degrees may be managed by specific functions [2-3].

In the field of artificial intelligence, neuro-fuzzy refers to combinations of artificial neural networks and fuzzy logic. Neuro-fuzzy was proposed by J. S. R. Jang. Neuro-fuzzy hybridization results in a hybrid intelligent system that synergizes these two techniques by combining the human-like reasoning style of fuzzy systems with the learning and connectionist structure of neural networks. Neuro-fuzzy hybridization is widely termed as Fuzzy Neural Network (FNN) or Neuro-Fuzzy System (NFS) in the literature [4]. Neuro-fuzzy system incorporates the human-like reasoning style of fuzzy systems through the use of fuzzy sets and a linguistic model consisting of a set of IF-THEN fuzzy rules. The main strength of neuro-fuzzy systems is that they are universal approximates with the ability to solicit interpretable IF-THEN rules [5-6].

Techniques for data classifications using Neuro-Fuzzy has been continually evolving to ensure efficient classification accuracy [6-7].

In this paper we have implemented the dynamic clustering algorithm and we have got each feature cluster separately. This output is forward to the transition process to reduce the size of cluster. Lastly the outputs of transition process are forward to the classification layer and classify the dataset. For classifying the dataset we have used the multi-perceptron classifier.

Using the above method it also combines the different class of element. Using this method we got the large number of cluster and when it create the cluster it also combine the different class of cluster. To reduce the cluster size we have implement the homogenous clustering process. It only combines the similar class of element.

2. PROPOSED WORK

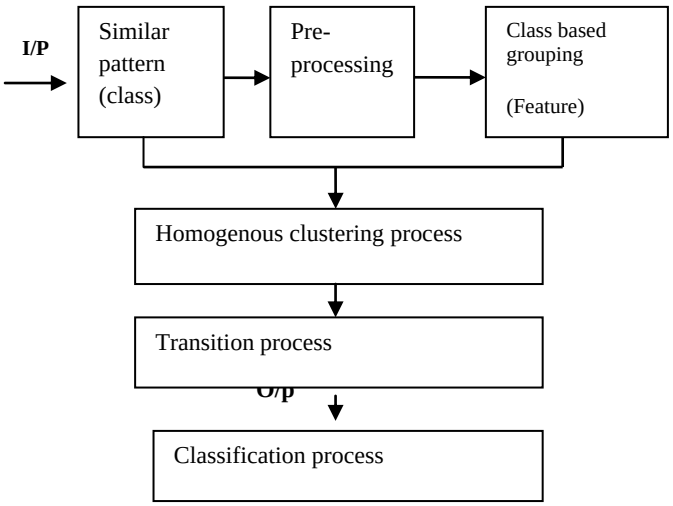

Fig 1:-. The Structure of the Homogeneous ClusteringProcess for Classifications Using Dynamic Clustering

The Fig. 1 shows the structure of dynamic clustering. The dynamic clustering method is divided into three key steps. Dynamic clustering process, transition process and classification process. The dynamic clustering process is consist two parts, the preprocessing of data and the class based grouping algorithm. In the preprocessing step the dataset is sorted in the ascending order and then class based grouping algorithm is performed. The dynamic clustering process is performed to find the number of membership functions proper to each feature. The second part is the transition process to renovate the original input data through a Gaussian membership function to develop a binary input. Lastly, the classification process is part of the learning and classification using neural network.

2.1 Algorithm Used: Class-Based grouping

Input:
xi: Input feature i.

Output:

Ni: Number of membership function of input feature i. ci: mean of each membership function of input feature i. σi : standard deviation of each membership function of input feature i.

Begin

Step1xi=Sort (xi);
j = 1

N = 1;

Step 2 and step 3 prevClass = Class of xij(similar class);

xij is a member of group 1;

For j = 2 to Number of record

If prevC ! = Cj Then

N = N + 1;

prevClass = Class of xij ;

End If

Step 4 xij is a member of group n;

End For

Step 5 threshold = Average of number of member of group;

Step6 Calculate centroid of each group; Repeat

Step 7 for j = 1 to N

If Number of group j < threshold Then Merge group j to other group that has nearest centroid and calculate new centroid;

N = N - 1;

End If

End For

Step 8 until No group that has a number of member <threshold

Step 9 Calculate c of each group;

Calculate I of each group;

Return (N, ci, i); End

The dynamic clustering algorithm also called as a "class based grouping" algorithm. The dynamic clustering algorithm works as follows [11].

The first step of algorithm to sort the element in ascending order and also sort with the similar class. The second step is group the element. If the previous element and next element class is same then group that element and say it cluster. The third step is group the cluster. The fourth step is finding the centroid value. The fifth step is finding membership values .The sixth step is finding the threshold value. The seven steps are again group that cluster using the threshold value .The eight steps is finding the centroid value. The ninth step is found out standard deviation.

3. RESULT AND DISCUSSION

3.1 Dynamic Clustering

In this paper the yeast dataset is used for calculating the result. The yeast dataset have the 9 different attributes and 10 different types of classes. The total size of dataset is 1484. Following are the attributes information.

Sequence Name(Accession number for the SWISS-PROT database), McGeoch's method for signal sequence recognition (mcg), von Heine's method for signal sequence recognition (gvh), Score of the ALOM membrane spanning region prediction program (alm), Score of discriminant analysis of the amino acid content of(mit), The N-terminal region (20 residues long) of mitochondrial and non-mitochondrial proteins, Presence of "HDEL" substring (thought to act as a signal for retention in the endoplasmic reticulum lumen). Binary attribute (erl), Peroxisomal targeting signal in the C-terminus (pox), Score of discriminant analysis of the amino acid content of vacuolar and extracellular proteins (vac), Score of discriminant analysis of nuclear localization signals of nuclear and non-nuclear proteins (nuc).

Following are the class information

Class is the localization site. Please see Nakai & Kanehisa referenced above for more details. Cytosolic or cytoskeletal (CYT), Nuclear(NUC), Mitochondrial membrane protein, no N-terminal signal (ME3), membrane protein, Uncleaved signal (ME2), membrane protein, cleaved signal (ME1), extracellular(EXC), vacuolar(VAC), peroxisomal (POX), Endoplasmic reticulum lumen (ERL).

In the following example the dynamic Clustering process is performed. In that process yeast dataset [UCI] is used. In the Table I, starts with storing the data point in features X_i in the increasing order. The above process is similar to the incremental clustering the first data point is assigned to the first cluster. In this step member is assigned to each group.

Table 1. Sorted values of feature 'mcg' for the yeast dataset

Sr. no.	Sorting the element	
	Xi(mcg)	***class***
1	0.40	CYT
2	0.40	CYT
3	0.42	MIT
4	0.42	NUC
5	0.43	MIT
6	0.43	NUC
7	0.43	CYT
8	0.43	CYT
9	0.45	CYT
10	0.46	CYT
11	0.46	CYT

Sr. no.	Sorting the element	
	Xi(mcg)	***class***
12	0.47	CYT
13	0.48	NUC
14	0.50	MIT
15	0.50	NUC
16	0.51	CYT

Table 2: class based formation of clusters for the 'mcg' feature values

Cluster	**Grouping the cluster**	***Cluster***	**Grouping the cluster**
	Member		***Member***
1	0.4 ,0.4	6	0.43,0.43,0.45,0.46,0.47
2	0.42	7	0.48
3	0.42	8	0.5
4	0.43	9	0.5
5	0.43	10	0.51
		11	0.51,0.58

In the table II groups are distributed by considering the cluster and its data point. Now group the values according to the class of each cluster. This table II is used for calculating the cluster centroid value and cluster standard deviation.

Table 3: Estimation of centroid & standard deviation of all clusters

Cluster	**Calculating centroid and S.D**	
	Centroid	***Standered deviation***
1	0.4	0.0
2	0.42	0.0
3	0.42	0.0
4	0.43	0.0
5	0.43	0.800
6	0.45	0.0
7	0.48	0.0
8	0.5	0.0
9	0.5	0.0
10	0.51	0.0
11	0.565	0.299

In the Table III all centroids of each cluster are calculated. In order to apply the agglomerative clustering each feature must have proper threshold, which can calculate from the equation (1) as following.

$$F_{th} = \frac{\sum_{j=1}^{N} M_j}{N} \qquad \text{.....(1)}$$

The *Fth* is the threshold of considered feature which is the average number of members in each cluster. The variable *Mj* is the number of members in cluster *j*. The *N* represents the total numbers of clusters derive from the first step. The *Fth* is used to judged existence of the cluster in the third step.

3.2 Transition Process

The number of clusters and initial values of the mean and standard deviation from a dynamic clustering process are used in the transition process. The original inputs are fed to the Gaussian membership function layer of the model. The number of membership functions for each feature is varies depending on the cluster of the feature [11].

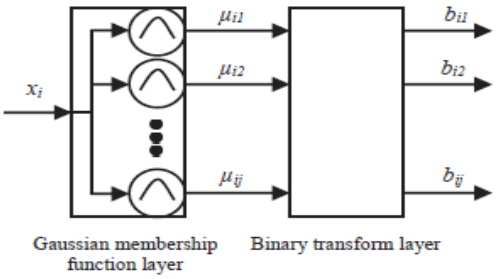

Fig 2: The structure of transition processes

The structure [11] in Fig.2 including Gaussian membership function layer and Binary transform layer. The membership values of feature *xi* are defined in equation (2).

$$\mu_{ij} \begin{cases} o, if \sigma_{ij} = 0 \; and \; x_i \neq c_{ij} \\ e^{\left(\frac{1}{2}\frac{(x_i - c_{ij})}{\sigma_{ij}^2}\right)^2}, if \; \sigma_{ij} \neq 0 \\ 1, if \; \sigma_{ij} = 0 \; and \; x_i = c_{ij} \end{cases} \qquad \text{...(2)}$$

Where *i* is identified the original feature and *j* is identified the cluster order of each feature. Normally the membership value calculated from Gaussian membership function. However the membership value is set to 0 if no distribution of that cluster and *xi* is not equal to mean of the cluster. The membership value is set to 1 if no distribution of that cluster and *xi* is equal to mean of the cluster [11].

Table 4. Result of transition process

Feature 1	**Gaussian value**
Cluster 1	[2.8423E-8,0.9795,0.4753,0.0820, 0.0820]
Cluster 2	[0.3678,0.3678]
Cluster 3	[1.0 , 1.0]
. . .	. . .
Cluster343	[0.223,1.0,0.22365]

The Table IV result shows the Gaussian value of 'msg' features cluster. These values are calculated from the eq (2).Using these Gaussian values we have forward to the classification process. To classify these dataset we have used the multi-perceptron classifier and classify the dataset .in these process we have got the 60 percent accuracy .These is the work of current paper. Now I have added the following step to improve the accuracy.

Following are the step to perform the homogeneous clustering process. In that we are combining the same class of element. After that we are performing the dynamic clustering algorithm. Using this homogenous clustering process I have found that it consume, the less time as comparison to other method and gives the best accuracy.

Following are steps and result of homogenous clustering.

Table V. Homogeneous clustering process:

0.49	0.39	0.52	0.29	0.50	0.00	0.48	0.22	EXC
0.81	0.85	0.47	0.37	0.50	0.00	0.56	0.22	EXC
0.80	0.63	0.45	0.29	0.50	0.00	0.51	0.22	EXC
0.78	0.74	0.42	0.26	0.50	0.00	0.43	0.22	ME1
0.75	0.70	0.38	0.27	0.50	0.00	0.49	0.22	ME1
0.44	0.38	0.48	0.32	0.32	0.50	0.83	0.53	POX
0.37	0.53	0.60	0.19	0.50	0.50	0.42	0.22	POX
0.43	0.67	0.48	0.27	0.50	0.00	0.53	0.22	MIT
0.64	0.62	0.49	0.15	0.50	0.00	0.53	0.22	MIT

In the homogenous clustering we are sorting the dataset in similar class wise and in ascending order. Using this homogenous cluster we have minimize the cluster size. Output of the homogenous clustering process is forward to the transition process. Lastly forward this output to the classification process.

Step 1: In the homogeneous clustering process we are sorting the dataset in similar class wise and in ascending order.

Using this method the disadvantages of our first step is overcome. Means it can't combine the different class.

Step 2: Now output of the first step is forward to the dynamic clustering algorithm. The dynamic clustering process perform three steps clustering, transition process and classification. The clustering steps create the cluster and calculating the mean and slandered deviation.

In these step we are creating the cluster feature wise direction and calculating the mean and standard deviation. Using these we have minimize the cluster size of each and every feature.

Table VI: Clusters Mean and SD

Feature 1	Clust no.	Mean	SD
	Cluster#1	0.16800	0.009839
	Cluster#2	0.198889	0.004536
	:	:	:
	Cluster#145	0.91700	0.18738
Feature 2	Cluster#1	0.135000	0.003536
	Cluster#2	0.175270	0.3354
	:	:	:
	Cluster#142	0.920000	0.011547
Feature 3	Cluster#1	0.260000	0.0
	:	:	:
	Cluster#108	0.713333	0.002981
Feature 4	Cluster#1	0.0	0.0
	:	:	:
	Cluster#130	0.865000	0.003536
Feature 5	Cluster#1	0.200000	0.0
	:	:	:
	Cluster#5	0.100000	0.0
Feature 6	Cluster#1	0.83000	0.0
	:	:	:
	Cluster#6	0.0	0.0
Feature 7	Cluster #1	0.85600	0.12300
	:	:	:
	Cluster 81	0..698750	0.013500
Feature 8	Cluster#1	0.110000	0.000000
	:	:	:
	Cluster#67		0.010104

Step 3: Perform the transition process. The transition process contains Gaussian membership function and binary function. We are performing the Gaussian membership function for the purpose of reducing the cluster size.

Step 4: Classify the dataset using the multi-perceptron neural network.

Accuracy 64.97035040431267

Mean Absolute Error 0.13888969473162796

Number of Instances 1484.0

4. CONCLUSION

In this paper we have implemented the dynamic clustering algorithm based on homogeneous clustering. In the original dynamic clustering it combines the different class of element. Which groups the features of same class in a cluster.in the dynamic clustering algorithm, features cluster are formed by balancing the number of members in each cluster, But in this, even the features of different class are grouped together in the

same cluster. To avoid this situation we have implemented the homogenous clustering. Then this homogenous clustering process is applied on dynamic clustering process and generates the result. Advantages of this homogenous clustering it minimize the cluster size of each feature and it also gives the impressive accuracy.

5. REFERENCES

[1] Cejas jesus "compensatory fuzzy logic",la habana revista de Ingenieria industrial .2011

[2] Valiant leslie "natures algorithms for learning and prospering in complex world ", New york 2013

[3] Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques," Morgan Kauf-mann, San Francisco, 2006. pp. 325-370.

[4] E. lughoter "envloing fuzzy system:methodologies, advanced concept and application".springer heidelberg 2011.

[5] S. S. Haykin, "Neural Networks and Learning Machines," 3rd Ed., Prentice Hall, New York, 2009.

[6] J. Han, M. Kamber and J. Pei, "Data Mining: Concepts and Techniques," Morgan Kauf-mann, San Francisco, 2006. pp. 325-370.

[7] K. Subramanian, R. Savitha and S. Suresh, "Complex-valued Networks(IJCNN), The 2012 International Joint Conference on. IEEE, 2012. pp. 1-7.

[8] T. Kondo, J. Ueno and S. Takao, "Hybrid feedback GMDH- type neural network self-selecting various neurons and its application to medical image diagnosis of lung cancer," In: Soft Computing and Intelligent Systems (SCIS) and 13th International Symposium on Advanced Intelligent Systems (ISIS), 2012 Joint 6th International Conference on. IEEE, 2012. pp. 1925-1930

[9] A. Gajate, R. E. Haber, P. I. Vega and J. R. "Alique A transductive neuro-fuzzy controller: application to a drilling process, Neural Networks, IEEE July 2010, Transactions , vol. 21, no. 7, pp. 1158-1167,.

[10] H. Rouabah, C. Abdelmoula and M. Masmoudi, "Behavior control of a mobile robot based on Fuzzy logic and Neuro Fuzzy approaches for monitoring wall," In: Design & Technology

[11] Narissara Eiamkanitchat and Poonarin Wongchomphu "Enhance Neuro-Fuzzy System for Classification using Dynamic Clustering", The 4th Joint International Conference on Information and Communication Technology Electronic and Electrical Engineering (JICTEE-2014).