Data Classification Using Combination of Five Machine Learning Techniques

Md. Habibur Rahman¹, Jesmin Akhter², Abu Sayed Md. Mostafizur Rahaman¹, Md. Imdadul Islam¹

¹Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh
²Institute of Information Technology, Jahangirnagar University, Dhaka, Bangladesh
Email: asmmr@juniv.edu

Abstract

Data clustering plays a vital role in object identification. In real life we mainly use the concept in biometric identification and object detection. In this paper we use Fuzzy Weighted Rules, Fuzzy Inference System (FIS), Fuzzy C-Mean clustering (FCM), Support Vector Machine (SVM) and Artificial Neural Network (ANN) to distinguish three types of Iris data called Iris-Setosa, Iris-Versicolor and Iris-Virginica. Each class in the data table is identified by four-dimensional vector, where vectors are used as the input variable called: Sepal Length (SL), Sepal Width (SW), Petal Length (PL) and Petal Width (PW). The combination of five machine learning methods provides above 98% accuracy of class identification.

Keywords

Co-Variance of Fuzzy Rule, Objective Function, Surface Plot, Confusion Matrix, Scatterplot and Accuracy of Detection

1. Introduction

In this paper five widely used methods: Fuzzy weighted rule, FIS, FCM, SVM and ANN are integrated in classification of Iris data. Several works related to the paper are mentioned in this section. In [1] authors use Adaptive Neuro-Fuzzy Inference System (ANFIS) and the Fuzzy Inference System (FIS) for professional blogger classification, where FIS provides better results compared to Classification Based on Associations (CBA). The combination of Artificial Neural Network (ANN) and ANFIS gives better classification, whereas the proposed ANFIS of the paper shows the best result which is 93%. The concept of FIS in data classification is also found in [2], where fault of electrical transmission line is de-
In [3], fuzzy weighted rules are used to classify Iris data using seven membership function (MFs). The average classification rate is found 96.48%, 96.06% and 96.7% for 7, 9 and 11 labels of MFs. The main drawback of the paper is that, it only deals with single method of classification; therefore we have the scope of inclusion of other data segregation algorithms. The fuzzy rule-based classification is found in [4] for classification of coronary artery disease data, where trapezoidal membership functions are used for input variables. The classification rate varies with different weighting rules, the maximum value is found 92.8% and that of minimum value is 71.8%. In this paper, we applied fuzzy c-mean clustering in Iris data classification; the similar concept is available in MR brain image segmentation in [5]. Here the entire algorithm of C-mean clustering is shown and the performance of image classification is compared with seven different methods and fuzzy c-mean clustering provides moderate result. Application of FCM in image classification is found in [6], where FCM is combined with Convolution Neural Network (CNN) to recognize tumors in the brain. The accuracy of detection is claimed by the auditors is 91%. Application of FCM is also found in image classification in [7][8]. The SVM in data classification is used in [9], where text based automatic task classification is done. The authors claim the accuracy of classification in the range of 82% to 99%. Similar concept is found in [10] for breast cancer diagnosis, where three different types of kernels are used and accuracy is found above 90% for all cases.

In this paper we combined all the five algorithms to classify Iris data, although the concept of the paper is applicable in any type of data or feature vector-based image classification. The main objective of the paper is to get high accuracy of data classification avoiding deep learning technique so that process time will remain low. Actually, inclusion of Fuzzy weighted rule plays a vital role in data classification. Most of the previous works did not include the Fuzzy weighted rule hence they have to include deep learning to acquire high accuracy of classification, which needs huge process time. The combination of five methods of the paper like [11] is found more robust compared to previous works. We compare the result of the paper (using same data set) with two previous works and found better result, which is shown in result section.

The rest of the paper is organized as: Section 2 provides theoretical analysis of five machine learning algorithms used in this paper for data classification, Section 3 provides results based on analysis of Section 2 and Section 4 concludes entire analysis.

2. Theory of Data Classification

2.1. Fuzzy Inference System (FIS)

Fuzzy Inference System (FIS) consists of three building blocks: Fuzzification, Inference and De-fuzzification. The numerical data is converted to Fuzzy symbols using membership functions (MFs) consisting of several variables, where
each variable has its range of numerical value. The above conversion technique is called Fuzzification. The Inference block deals with some rules using if-then form to relate input and output. Finally output symbols are converted to numerical value using De-fuzzification technique on the output MFs.

2.2. Fuzzy Weighted Rule

The detail analysis of Fuzzy weighted rule is shown in [3] with numerical example. In this paper we show the steps of the algorithm in a different way like below:

**Algorithm 1: Fuzzy Weighted Rule**

1. Take the numerical data in tabular form as the input, where the size of each record is \( N \) and number of output types are \( n \)
2. Take \( M \) number of MFs for each of \( N \) input variable
3. Convert numerical input data of the table into Fuzzy linguistic values using MFs
4. Take union of Fuzzy linguistic value of 1st field, 2nd field, 3rd field, …, \( N \)th field for the case of first output
5. If the sets obtained from the unions of step 4 are: \( \{S_1\}, \{S_2\}, \{S_3\}, \ldots, \{S_N\} \) then the \( N \)-tuple \( ((\{S_1\}, \{S_2\}, \{S_3\}, \ldots, \{S_N\})\), First output) known as \( \text{R}_{\text{First output}} \)
6. Repeat step 4 and 5 to get \( \text{R}_{\text{Second output}}, \text{R}_{\text{Third output}}, \ldots, \text{R}_{\text{Nth output}} \)
7. Take union of 1st element of \( \text{R}_{\text{First output}}, \text{R}_{\text{Second output}}, \text{R}_{\text{Third output}}, \ldots, \text{R}_{\text{Nth output}} \) to get the rule \( \text{R}_k \)
8. Repeat step 7 for 2nd, 3rd, …, \( N \)th elements of \( \text{R}_{\text{First output}}, \text{R}_{\text{Second output}}, \text{R}_{\text{Third output}}, \ldots, \text{R}_{\text{Nth output}} \) to get the rule \( \text{R}_k, \text{R}_l, \text{R}_m, \ldots, \text{R}_N \)
9. Take the sum of non-overlapping range and full range of first input variable against all the \( n \) output
10. Take the ratio \( v_1 \) two terms of step 9
11. Repeat step 9 and 10 for the rest of input variables
12. Take \( \max(v_1,v_2,v_3,\ldots,v_N) \) and weights, \( W_j = \frac{V_j}{\max(V_1,V_2,\ldots,V_N)} \), where \( i = 1,2,3,\ldots,N \)
13. For each input record of \( N \)-tuple determine weighted co-variance of each rule like,
\[
R = \sum_{i=1}^{N} \Psi_{i,j}(X_j)W_j
\]
14. The highest value of \( R \) corresponding to \( k \)-th rule indicates the input tuple is under the output of \( k \)-th category

In this subsection few numerical examples are shown according to the steps Fuzzu weighted rule. First of all, we take few data of Iris under three categories.
called: Iris-Setosa, Iris-Versicolor and Iris-Virginica shown in Table 1. For each category four types of inputs (SL, SW, PL and PW) and corresponding output are taken as the initial data shown in Table 1. For better understanding of reader, we chose the same initial data of [3] and we elaborate the initial data processing steps more explicitly compared to previous paper.

For each input SL, SW, PL or PW we consider 7 trapezoidal membership functions named: HN, MN, SN, Z, SP, MP and HP as shown in Figures 1(a)-(d) for four input variables. The MFs of three output classes is shown in Figure 2.

2.3. Fuzzy c-Means Clustering

The main objective of FCM is to minimize the objective function,

\[ J_m = \sum_{j=1}^{c} \sum_{x(i) \in c_j} u_{ij}^m \left( \|x(i) - c_j\|^2 \right) \]  

(3)

where

- \( m \) is a real number greater than 1 called fuzzifier
- \( u_{ij} \) is the degree to which an \( x(j) \) belongs to the cluster \( j \) with center \( c_j \)
- \( x(j) \) is the \( j \)th data point
- \( c \) is the number of clusters

The steps of Fuzzy c mean clustering algorithm is given below like [12] [13].

**Algorithm 2: Fuzzy c-means clustering**

1. First consider \( n \) data points, \( x = [x(1) \ x(2) \ x(3) \ \ldots \ x(n)] \) to be segregated into \( c \) clusters
2. Take the initial value of center of clusters, \( c_k \) where \( k = 1, 2, 3, \ldots, c \)
   Evaluate grade (or degree) of membership \( u_{ij} \) i.e. the degree to which an \( x(j) \) belongs to the cluster with center \( c_j \),
   \[ u_{ij} = \frac{1}{\sum_{l=1}^{c} \left( \frac{x(i) - c_l}{x(i) - c_j} \right)^{1/m}} ; \quad \text{for} \quad i = 1, 2, 3, \ldots, n \]  
   (4)
3. The entire vector is expressed at \( j \)th iteration as,
   \( U_j(k) = [u_{11}, u_{21}, u_{31}, \ldots, u_{nk}] \) , under cluster 1
   \( U_j(k) = [u_{12}, u_{22}, u_{32}, \ldots, u_{nk}] \) , under cluster 2
   : 
   \( U_j(k) = [u_{1c}, u_{2c}, u_{3c}, \ldots, u_{nc}] \) , under cluster \( c \)
   Update the center \( c_j \) like,
   \[ c_j = \frac{\sum_{x(i) \in c_j} u_{ij}^m x(i)}{\sum_{x(i) \in c_j} u_{ij}^m} \]  
   (5)
4. Repeat step 3 and 4 until \( \left| U_j(k) - U_j(k+1) \right| < \varepsilon \) \( j = 1, 2, 3, \ldots, c \)

2.4. Support Vector Machine

The SVM is a supervised learning algorithm used for data classification,
Figure 1. MFs of four input variables. (a) MFs of input SL; (b) MFs of input PW; (c) MFs of input PL; (d) MFs of input SW.
Table 1. Three types of Iris data [3].

|       | SL  | SW  | PL  | PW  | Out |
|-------|-----|-----|-----|-----|-----|
| Iris-Setosa | 4.6 | 3.4 | 1.4 | 0.3 | 1   |
|         | 5.7 | 3.8 | 1.7 | 0.3 | 1   |
|         | 5.2 | 3.4 | 1.4 | 0.2 | 1   |
|         | 4.5 | 2.3 | 1.3 | 0.3 | 1   |
|         | 4.4 | 3.2 | 1.3 | 0.2 | 1   |
| Iris-Virginica | 6.1 | 3.0 | 4.9 | 1.8 | 3   |
|         | 6.1 | 2.6 | 5.6 | 1.4 | 3   |
|         | 6.9 | 3.1 | 5.4 | 2.1 | 3   |
|         | 6.7 | 3.1 | 5.4 | 2.4 | 3   |
|         | 6.2 | 3.4 | 5.4 | 2.3 | 3   |
| Iris-Versicolor | 6.6 | 2.9 | 4.6 | 1.3 | 2   |
|         | 5.0 | 2.0 | 3.5 | 1.1 | 2   |
|         | 6.2 | 2.2 | 4.5 | 1.5 | 2   |
|         | 5.9 | 3.2 | 4.8 | 1.8 | 2   |
|         | 6.1 | 2.9 | 4.5 | 1.5 | 2   |

Figure 2. MFs of three output classes.

decision-making, pattern recognition, forecasting of data, disease diagnostic etc. The SVM algorithm classifies objects taking decision boundary called hyperplane where the optimum hyperplane separates the points corresponding to objects with widest margin as discussed in [14] [15]. The generalized equation of a hyperplane like,

\[ f(x) = b + w^T x; \]

where \( w \) is known as the weight vector and \( b \) as the bias.

The SVM determines the constants: \( w^T, b, \tau \) such that \( w^T x + b \geq \tau \) for one group of points, \( w^T x + b \leq \tau \) for another group of points. The SVM uses Kernel function to provide the best trajectory of decision boundary.

2.5. Artificial Neural Network

In this paper we used feed-forward ANN, where signal only travels in one direc-
tion i.e. from input to output. Such neural network is called multi-layer perceptron and used for pattern recognition. We used it for the case of 10 and 20 hidden layers to observe relative performance. We also used ANN under backpropagation algorithm, where signal flows in both directions. The concept of both of above ANN is available in [16] [17] and here we avoid the theoretical analysis of such ANNs.

The five machine learning methods will be combined using Shannon entropy-based algorithm.

3. Result and Discussion

This section provides results based on theoretical analysis of previous section. First of all, we apply FIS on the Iris data. The FIS used in this paper is shown in Figure 3, where 7 MFs are used against each of the four input variables. We apply 69 Fuzzy rules and few of them are shown in Figure 4. The surface plot variables: PS, PL, PW and SW of the FIS is shown in Figures 5(a)-(f). Here the surface level 1.5, 2 and 2.5 provides the results of Iris-Set, Iris-Ver and Iris-Vir respectively. Next, we apply Fuzzy weighted rule on 150 data of Iris. The detail of the Fuzzy weighted rule is shown in Section 2.1. We run the algorithm 5 times taking 100 data each time, corresponding accuracy of correct recognition is given in Table 2 at the end of this section.

![Figure 3. Fuzzy system of data classification.](image)

![Figure 4. Some fuzzy rules.](image)
Next we apply Fuzzy c-mean clustering on the entire dataset taking two variables at a time. The scatterplot of three output data are shown in Figure 6. Few data points seem to cross its region i.e. produce some recognition error. Here 50 data for Iris-Set, 50 data for Iris-Ver and 50 data for Iris-Vir are taken.

Figure 5. Surface plot of the FIS. (a) Surface plot of PW vs. SL; (b) Surface plot of PL vs. SL; (c) Surface plot of SW vs. SL; (d) Surface plot of PL vs. PW; (e) Surface plot of SW vs. PW; (f) Surface plot of SW vs. PL.
Finally, scatterplot of data points in four combinations of four input variables are shown in Figures 7(a)-(d) to get the idea of best separation case. Here PW vs. PL shows the best separation as found in Figure 6(b). The regional separation of data points using SVM is shown in Figures 8(a)-(d), where Figure 8(b) shows the best regional separation. In future we will apply multiple linear regression (MLR) on four-dimensional input data to convert them into two-dimensional data, then apply SVM to observe any improvement compared to four cases of Figure 8.

Next, Irish data classification is done using feedforward ANN. The performance of the network, error histogram and confusion matrix are shown in Figure 9-11 for the case of 10 and 20 hidden layers. Similar results are shown in Figure 12 and Figure 13 for backpropagation ANN for 8 and 10 hidden layers. The performance is found better with increment of hidden layer at the expense of process time.

Except Weighted Fuzzy, no individual method provides high accuracy of recognition visualized from Table 2. The Weighted Fuzzy provides high accuracy at

| Experiments | Weighted Fuzzy | FIS | Fuzzy C-mean | SVM | Feedforward ANN | Backpropagation ANN | Combined |
|-------------|---------------|-----|--------------|-----|-----------------|---------------------|----------|
| 1           | 0.931         | 0.881 | 0.892       | 0.873 | 0.835       | 0.878 | 0.974 |
| 2           | 0.904         | 0.855 | 0.879       | 0.907 | 0.862       | 0.874 | 0.982 |
| 3           | 0.929         | 0.867 | 0.862       | 0.893 | 0.857       | 0.895 | 0.988 |
| 4           | 0.913         | 0.853 | 0.864       | 0.880 | 0.866       | 0.903 | 0.976 |
| 5           | 0.932         | 0.871 | 0.882       | 0.921 | 0.841       | 0.917 | 0.978 |

**Figure 6.** Scatterplot of data under fuzzy c-mean clustering.
**Figure 7.** Scatterplot of Iris data.
Figure 8. Three regions of output under SVM.

Figure 9. Performance of the feedforward ANN. (a) 10 hidden neuron; (b) 20 hidden neuron.

Figure 10. Error histogram. (a) 10 hidden neurons; (b) 20 hidden neurons.
the expense of process time, but process time is much smaller than deep leaning technique. We combined five methods using entropy based combining algorithm of [11], which provides accuracy of recognition above 98% for all the five experiments. Finally, we compared our results with NN + SVM of [18] and FCM + SVM of [19], using the same data, where the result of first case is found 0.9417 and that of second case is 0.9445. Our model is the combination of five MLs, which is more robust than previous works in data classifications.
Figure 13. Error histogram of backpropagation ANN. (a) 8 hidden neuron; (b) 10 hidden neuron.

4. Conclusion

In this paper Iris data classification is done using FIS, Weighted Fuzzy rule, Fuzzy c-mean clustering, SVM and ANN. The combination of five techniques gives minimum value of accuracy of 97.4%, which is found better than previous individual method. The concept of the research work is also applicable for any type of tabular data. The high accuracy of classification of the paper is found because of inclusion of weighted fuzzy rule. The process time of weighted fuzzy rule is larger than the other five techniques used in the paper but considerably lower than deep learning like Convolutional Neural Network (CNN). The proposed technique of the paper provides high accuracy with minimum possible process time. Still we have scope to include other machine learning techniques like: Principal Component Analysis, Linear Discriminant Analysis (LDA), Bayesian Classification, Decision tree etc.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Asim, Y., Raza, B., Malik, A.K., Shahid, A.R., Faheem, M. and Kumar, Y.J. (2019) A Hybrid Adaptive Neuro-Fuzzy Inference System (ANFIS) Approach for Professional Bloggers Classification. 2019 22nd International Multitopic Conference (INMIC), Islamabad, 29-30 November 2019, 1-6. https://doi.org/10.1109/INMIC48123.2019.9022776

[2] Bhatnagar, M. and Yadav, A. (2020) Fault Detection and Classification in Transmission Line Using Fuzzy Inference System. 2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE), Jaipur, 1-3 December 2020, 1-6. https://doi.org/10.1109/ICRAIE51050.2020.9358386

DOI: 10.4236/jcc.2021.912004
[3] Chen, Y.-C., Wang, L.-H. and Chen, S.-M. (2006) Generating Weighted Fuzzy Rules from Training Data for Dealing with the Iris Data Classification Problem. *International Journal of Applied Science and Engineering*, 4, 41-52.

[4] Mohammadpour, R.A., Abedi, S.M., Bagheri, S. and Ghaemian, A. (2015) Fuzzy Rule-Based Classification System for Assessing Coronary Artery Disease. *Computational and Mathematical Methods in Medicine*, 2015, Article ID: 564867. https://doi.org/10.1155/2015/564867

[5] Meena Prakash, R. and Shantha Selva Kumari, R. (2016) Fuzzy C Means Integrated with Spatial Information and Contrast Enhancement for Segmentation of MR Brain Images. *International Journal of Imaging Systems and Technology*, 26, 116-123. https://doi.org/10.1002/ima.22166

[6] Rao, L.J., Challa, R., Sudarsa, D., Naresh, C. and Basha, C.Z. (2020) Enhanced Automatic Classification of Brain Tumours with FCM and Convolution Neural Network. 2020 3rd International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, 20-22 August 2020, 1233-1237. https://doi.org/10.1109/ICSSIT48917.2020.9214199

[7] Shang, R., Xie, K., Okoth, M.A. and Jiao, L. (2019) SAR Image Change Detection Based on Mean Shift Pre-Classification and Fuzzy C-Means. 2019 *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* 2019, Yokohama, 28 July-2 August 2019, 2358-2361. https://doi.org/10.1109/IGARSS.2019.8898464

[8] Mantilla, L. (2019) Classification of Satellite Images Using Rp Fuzzy C Means for Unsupervised Classification Algorithm. 2019 *IEEE Colombian Conference on Applications in Computational Intelligence (ColCACI)*, 5-7 June 2019, Barranquilla, 1-5. https://doi.org/10.1109/ColCACI.2019.8781988

[9] Shin, H. and Paek, J. (2018) Automatic Task Classification via Support Vector Machine and Crowdsourcing. *Mobile Information Systems*, 2018, Article ID: 6920679. https://doi.org/10.1155/2018/6920679

[10] Wang, H., Zheng, B., Yoon, S.W. and Ko, H.S. (2018) A Support Vector Machine-Based Ensemble Algorithm for Breast Cancer Diagnosis. *European Journal of Operational Research*, 267, 687-699. https://doi.org/10.1016/j.ejor.2017.12.001

[11] Azad, M.A.K., Majumder, A., Das, J.K. and Islam, M.I. (2021) Improving Signal Detection Accuracy at FC of a CRN Using Machine Learning and Fuzzy Rules. *Indonesian Journal of Electrical Engineering and Computer Science*, 21, 1140-1150. https://doi.org/10.11591/ijeecs.v21.i2.pp1140-1150

[12] Yu, T., Yang, J. and Lu, W. (2019) Dynamic Background Subtraction Using Histograms Based on Fuzzy C-Means Clustering and Fuzzy Nearness Degree. *IEEE Access*, 7, 14671-14679. https://doi.org/10.1109/ACCESS.2019.2893771

[13] Sivanagaleela, B. and Rajesh, S. (2019) Crime Analysis and Prediction Using Fuzzy C-Means Algorithm. 2019 3rd *International Conference on Trends in Electronics and Informatics (ICOEI)*, Tirunelveli, 23-25 April 2019, 595-599. https://doi.org/10.1109/ICOEI.2019.8862691

[14] Duan, M. (2018) Short-Time Prediction of Traffic Flow Based on PSO Optimized SVM. 2018 *International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)*, Xiamen, 25-26 January 2018, 41-45. https://doi.org/10.1109/ICITBS.2018.00018

[15] Mu, W., Zou, Z., Sun, H., Liu, G., Xia, G. and Wang, S. (2018) Model Classification of Guided Wave Signal Based on the Visibility Graph and SVM. 2018 *IEEE Far East NDT New Technology & Application Forum (FENDT)*, Xiamen, 6-8 July 2018, 156-160. https://doi.org/10.1109/FENDT.2018.8681982
[16] Saxena, U., Kaushik, D., Bansal, M., Chandel, H., Sahu, U. and Bhowmik, D. (2018) Low Energy Implementation of Feedforward Neural Network with Backpropagation Algorithm Using a Spin Orbit Torque Driven Skyrmionic Device. 2018 IEEE International Magnetics Conference (INTERMAG), Singapore, 23-27 April 2018, 1. https://doi.org/10.1109/INTMAG.2018.8508069

[17] Lu, J., Luo, X., Liu, D., Liu, P. and Liu, B. (2019) A Configurable Architecture of ANN in Hardware with Resource-Efficient Reusable Neuron. 2019 IEEE 13th International Conference on ASIC (ASICON), Chongqing, 29 October-1 November 2019, 1-4. https://doi.org/10.1109/ASICON47005.2019.8983505

[18] Liu, H., Xiao, X., Li, Y., Mi, Q. and Yang, Z. (2019) Effective Data Classification via Combining Neural Networks and SVM. 2019 Chinese Control and Decision Conference (CCDC), Nanchang, 3-5 June 2019, 4006-4009. https://doi.org/10.1109/CCDC.2019.8832442

[19] Saini, R., Kumar, P., Roy, P.P. and Dogra, D.P. (2017) An Efficient Approach for Trajectory Classification Using FCM and SVM. 2017 IEEE Region 10 Symposium (TENSYMP), Cochin, 14-16 July 2017, 1-4. https://doi.org/10.1109/TENCONSpring.2017.8070076