Scientific Research Publishing

Journal of Computer and Communications, 2020, 8, 79-89
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Data Prediction Model Using Combination of Clustering and Fuzzy Technique

Md. Mafiul Hasan Matin, Tanzim Kabir, Amina Khatun, Md. Imdadul Islam

Department of Computer Science and Engineering, Jahangirnagar University, Dhaka, Bangladesh
Email: mafiulmatinju@gmail.com, tanzimkabir29@gmail.com, amina_bashar@yahoo.com, imdad@juniv.edu

How to cite this paper: Matin, Md.M.H., Kabir, T., Khatun, A. and Islam, Md.I. (2020) Data Prediction Model Using Combination of Clustering and Fuzzy Technique. *Journal of Computer and Communications*, **8**, 79-89.
https://doi.org/10.4236/jcc.2020.87007

Received: June 11, 2020
Accepted: July 27, 2020
Published: July 30, 2020

Abstract

The analysis of environmental daily evaporation plays a vital role in the field of agriculture. It is very essential to know the daily evaporation rate of a particular area for proper cultivation. So, we need a standard prediction model which can predict the daily evaporation. In this paper, we use subtractive clustering and Fuzzy logic to predict daily evaporation of a particular area. The input data used in the paper are: maximum soil temperature, average soil temperature, average air temperature, minimum relative humidity, average relative humidity and total wind, which are related to the daily evaporation of a particular area as the output. The accuracy of output of the paper is compared with the previous model of Artificial Neural Network (ANN) and we get better result towards the target value. The finding of the paper is applicable in environmental science, geological science and agriculture.

Keywords

Subtractive Clustering, Fuzzy Interface System, ANN, Scatterplot and Surface Plot

1. Introduction

Future data prediction (short or long term forecasting) is essential in any engineering design, which is done using different machine learning algorithms. This section deals with some state-of-art works in data prediction. Application of Fuzzy system and Neural Network (NN) is found in GDP forecasting of IRAN in [1]. The prediction accuracy of Neural-Fuzzy is found 5.92% and that of Fuzzy-logic is 6.46%. In [2], T-S fuzzy neural network prediction model is used in prediction of the photovoltaic power generation in short term basis. The prediction results of the paper are compared with traditional back-propagation (BP) neural network. The average relative error of proposed method of the paper is found 5.61% and that of BP is 10.43%. Similar analysis is found in [3], where hybrid radial basis function (RBF) neural network algorithm is used for forecasting road speed. The mean absolute percentage error (MAPE) of the proposed method was found minimum comparing with other three methods: time series method, BP neural network and RBF neural network. Another application of Fuzzy Neural Network using Improved Decision Tree is found in [4], where short term electrical load forecasting is done. The predicted load and predictive error of the proposed model is found better compared to few previous models. In [5] the author used a combination of Artificial Neural Network and Fuzzy logic to analyze the strength of cement. The authors compared the results with a pure ANN model. Upon doing so they found their own model to be more user-friendly and easier to use, rather than a pure ANN. Data prediction models are also found in biomedical applications, for example in [6], ANFIS (adaptive neuro-fuzzy inference system) is applied in prediction of epilepsy, analyzing electroencephalogram (EEG) signals. The paper reveals MSE of both training data and test data for 9 patients. In [7] combination of Fuzzy C clustering and BP neural network is used in short-term electricity consumption forecasting. Authors consider the following input parameters: maximum temperature, minimum temperature, maximum humidity, minimum humidity, wind power and air quality with different weights. The average error of BP neural network forecasting method is found 15.44% and that of proposed algorithm is 6.94%. Sometimes financial data are also predicted by ANN, for example in [8] authors predict the revenue based on previous data. The MSE is varied taking number of hidden neurons and best performance epoch as parameters. Another example of financial data analysis is found in [9], where authors carry out stock market prediction using supervised machine learning. In this paper we use subtractive clustering and Fuzzy Inference System (FIS) to predict "environmental daily evaporation" which is useful for cultivation of crops anywhere in the world.

The remaining portion of the paper is arranged as follows: Section 2 provides the basic theory of fuzzy system and clustering techniques, Section 3 presents the system model used to derive the target output, Section 4 provides results based on analysis of Section 3 and finally, Section 5 concludes the entire analysis.

2. Basic Theory of Fuzzy System and Clustering Techniques

To solve Engineering problem we use Fuzzy Inference System (FIS) consists of three parts: Fuzzification, Inference (rule based) and De-fuzzification like **Figure 1**. Fuzzification is the process of converting a crisp input value or conventional numerical data to a Fuzzy value based on our knowledge or using grade of membership function (MF). De-fuzzification is the reversed way *i.e.* Fuzzy to crisp conversion. Fuzzy inference is the process of formulating the mapping from a given input to an output using fuzzy logic.

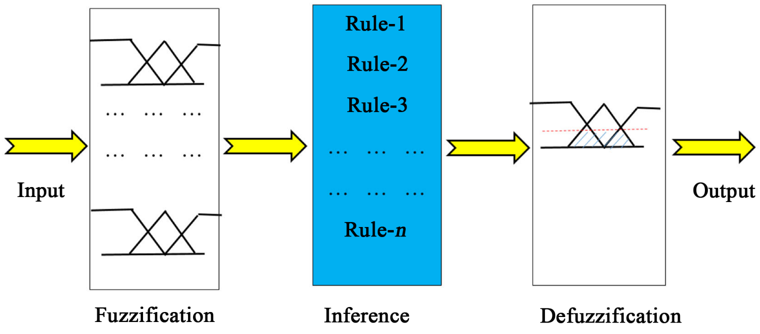


Figure 1. Basic architecture of FIS.

Next part of the section deals with basic theory of subtractive clustering algorithm. Basic task of data clustering is to provide several groups or clusters from the whole data set, where data or objects under a cluster are more alike compared to data or object of other clusters. Different types of clustering algorithms: Fuzzy C-mean clustering, K-mean clustering, subtractive clustering, Gaussian (EM) clustering algorithm, etc. are used in the field of pattern recognition, image or signal analysis, information retrieval, bioinformatics, machine learning, etc. In this paper we use a basic subtractive clustering algorithm since it can be integrated with FIS quite easily.

Subtractive clustering algorithm starts by assuming each data point in the dataset to be a potential centroid. Let us consider the data set, $S=\{x_1,x_2,x_3,\cdots,x_N\}$ and a neighborhood radius r_a. Now we have to evaluate the density of surrounding data points about each point x_i. The point having the maximum number of adjacent points within the radius r_a is selected as the first centroid. The density of surrounding points around x_i is evaluated by a parameter called mountain function $D(x_i)$, expressed as [10] [11],

$$D(x_i)=\sum_{j=1}^{N}\exp\left(-\frac{\|x_i-x_j\|^2}{\left(\frac{r_a}{2}\right)^2}\right) \tag{1}$$

where $\|x_i-x_j\|^2$ is the Euclidian distance between x_i and x_j. We have to select the data point $x_i{}^*$ whose mountain function provides the maximum value. The value of $D(x_i)$ increases with increase in the number of neighbors x_j around the data point x_i and also increases if the distances of neighboring points reduce.

The magnitude of mountain function of a potential centroid is reduced heavily if it is located near the first centroid. The expression of mountain function is modified to eliminate the impact of first centroid like [12] [13],

$$D_k(x_i)=D_{k-1}(x_i)-D_{cen}\sum_{j=1}^{N}\exp\left(-\frac{\|x_j-x_{cen}\|^2}{\left(\frac{r_b}{2}\right)^2}\right) \tag{2}$$

Here, x_{cen} is the first centroid, D_{cen} is the density value of the first centroid,

$D_{k-1}(x_i)$ is the mountain function of previous step and typically $r_b = 1.5r_a$. The increment of radius $r_b = 1.5r_a$ causes the density value (magnitude of mountain) of data points near the first centroid to be lower compared to data points that are farther away. This promotes the creation of clusters whose centroids are farther away from each other. Similarly, the data point with greatest density value achieved by (2) will be selected as the second centroid. Above process is continued until getting density value greater than a threshold and the remaining data points are considered as ordinary points on the scatterplot.

3. System Model

In this paper, our main objective is to combine two techniques: subtractive clustering and Fuzzy system to predict better output compared to individual one as shown in Figure 2. We consider input data of six environmental variables: Maximum soil temperature, Average soil temperature, Average air temperature, Minimum relative humidity, Average relative humidity and Total wind. The prediction output of the system is daily evaporation, which is another environmental variable. The training data used in the system of the paper is shown in appendix (**Table 1**).

The dataset is first fed to a subtractive clustering system to generate natural clusters among the data. The clustered data is then imported into the Fuzzy Inference System (FIS), where the model is trained. After the model is trained, it accepts input variable values from the user and can successfully predict the output with considerable accuracy. The FIS model is used to relate the six input variable with the out shown in Figure 3(a), where each input variable possesses seven membership functions (each cluster of subtractive clustering correspond to its MF) shown in Figure 3(b). In this paper we use Gaussian MF and in practical FIS, the MFs are not uniform like Figure 3(b), their width and phase shift are non-uniform. The basic structure of the FIS is shown in Figure 3(c). Few Fuzzy rules are shown in Figure 4.

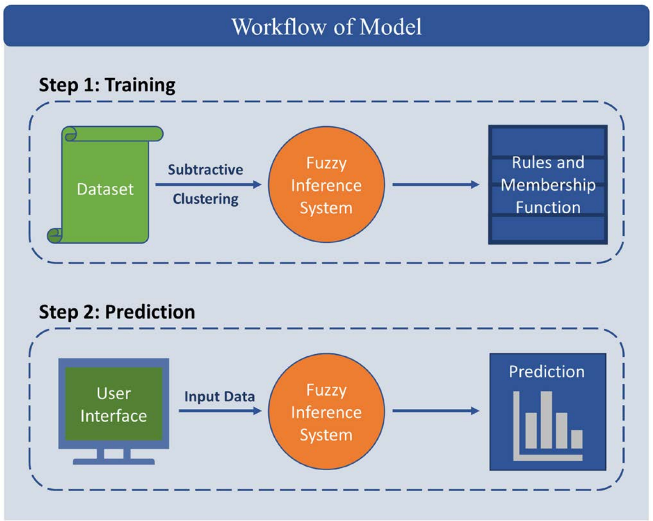
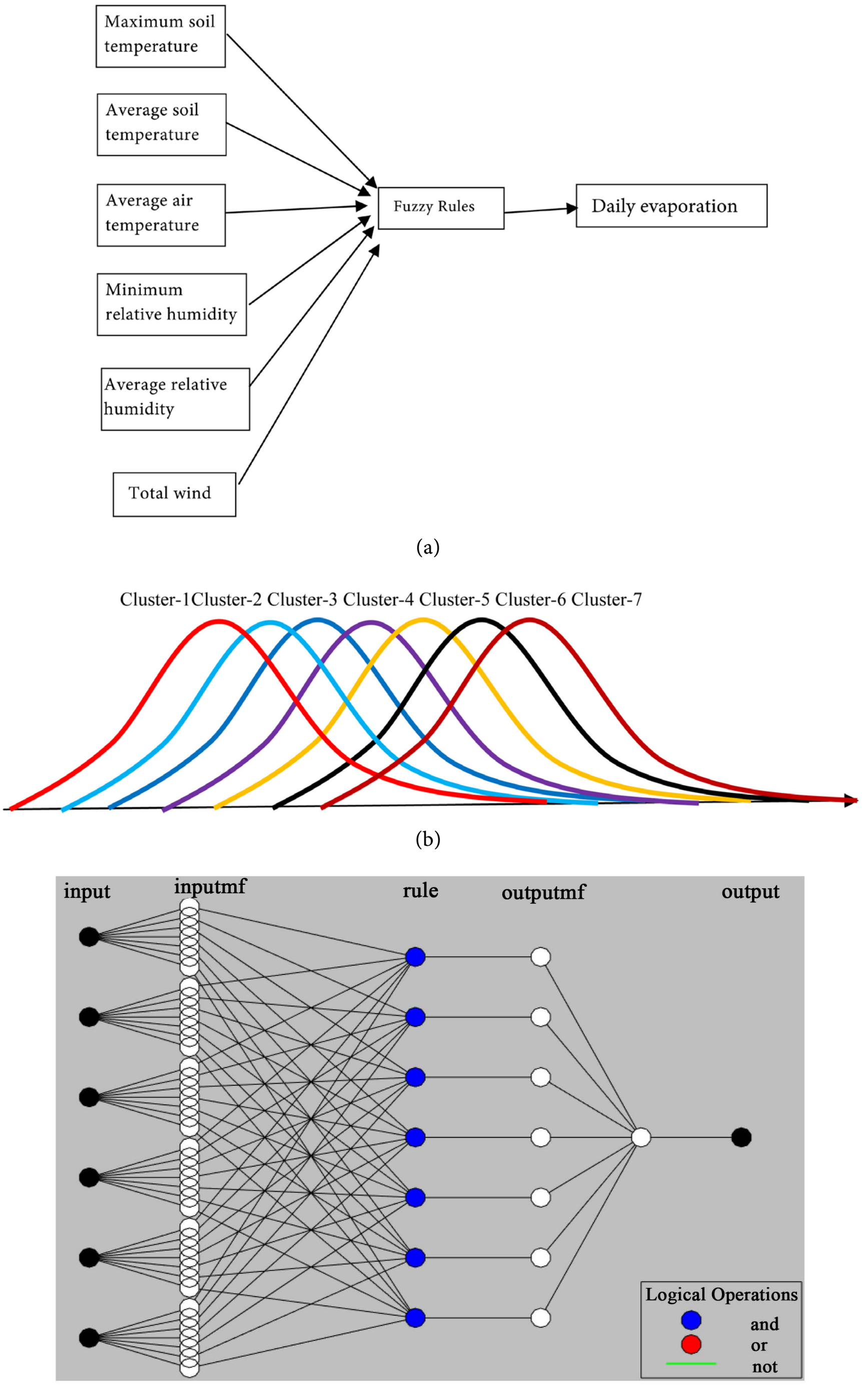


Figure 2. Workflow of proposed model.

Table 1. Training data.

Max Soil Temp	Avg Soil Temp	Avg Air Temp	Min Rel Hum	Avg Rel Hum	Total Wind	Daily Evap
84	147	151	40	398	273	30
84	149	159	28	345	140	34
79	142	152	41	388	318	33
81	147	158	50	406	282	26
84	167	180	46	379	311	41
74	131	147	73	478	446	4
73	131	159	72	462	294	5
75	134	159	70	464	313	20
84	161	195	63	430	455	31
86	169	206	56	406	604	38
88	178	208	55	393	610	43
90	187	211	51	385	520	47
88	171	211	54	405	663	45
88	171	201	51	392	467	45
81	154	167	61	448	184	11
79	149	162	59	436	177	10
84	160	173	42	392	173	30
84	160	177	44	392	76	29
84	168	169	48	398	72	23
77	147	170	60	431	183	16
87	166	196	44	379	76	37
89	171	199	48	393	230	50
89	180	204	48	394	193	36
93	186	201	47	386	400	54
93	188	206	47	389	339	44
94	199	208	45	370	172	41
93	193	214	50	396	238	45
93	196	210	45	380	118	42
96	198	207	40	365	93	50
95	202	202	39	357	269	48
84	173	173	58	418	128	17
91	170	168	44	420	423	20
88	179	189	50	399	415	15
89	179	210	46	389	300	42
91	182	208	43	384	193	44
92	196	215	46	389	195	41
94	192	198	36	380	215	49
96	195	196	24	354	185	53

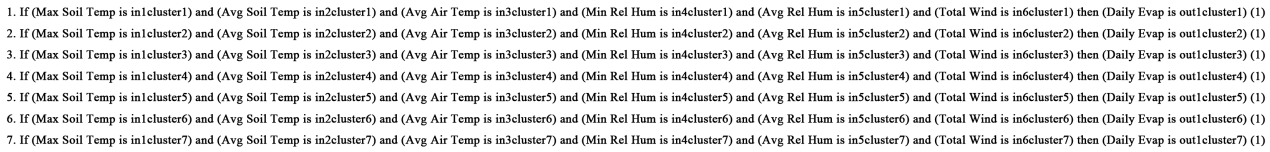


(a)

(b)

(c)

Figure 3. The FIS of the proposed model. (a) FIS model; (b) Seven MFs of seven clusters; (c) FIS structure.



1. If (Max Soil Temp is in1cluster1) and (Avg Soil Temp is in2cluster1) and (Avg Air Temp is in3cluster1) and (Min Rel Hum is in4cluster1) and (Avg Rel Hum is in5cluster1) and (Total Wind is in6cluster1) then (Daily Evap is out1cluster1) (1)
2. If (Max Soil Temp is in1cluster2) and (Avg Soil Temp is in2cluster2) and (Avg Air Temp is in3cluster2) and (Min Rel Hum is in4cluster2) and (Avg Rel Hum is in5cluster2) and (Total Wind is in6cluster2) then (Daily Evap is out1cluster2) (1)
3. If (Max Soil Temp is in1cluster3) and (Avg Soil Temp is in2cluster3) and (Avg Air Temp is in3cluster3) and (Min Rel Hum is in4cluster3) and (Avg Rel Hum is in5cluster3) and (Total Wind is in6cluster3) then (Daily Evap is out1cluster3) (1)
4. If (Max Soil Temp is in1cluster4) and (Avg Soil Temp is in2cluster4) and (Avg Air Temp is in3cluster4) and (Min Rel Hum is in4cluster4) and (Avg Rel Hum is in5cluster4) and (Total Wind is in6cluster4) then (Daily Evap is out1cluster4) (1)
5. If (Max Soil Temp is in1cluster5) and (Avg Soil Temp is in2cluster5) and (Avg Air Temp is in3cluster5) and (Min Rel Hum is in4cluster5) and (Avg Rel Hum is in5cluster5) and (Total Wind is in6cluster5) then (Daily Evap is out1cluster5) (1)
6. If (Max Soil Temp is in1cluster6) and (Avg Soil Temp is in2cluster6) and (Avg Air Temp is in3cluster6) and (Min Rel Hum is in4cluster6) and (Avg Rel Hum is in5cluster6) and (Total Wind is in6cluster6) then (Daily Evap is out1cluster6) (1)
7. If (Max Soil Temp is in1cluster7) and (Avg Soil Temp is in2cluster7) and (Avg Air Temp is in3cluster7) and (Min Rel Hum is in4cluster7) and (Avg Rel Hum is in5cluster7) and (Total Wind is in6cluster7) then (Daily Evap is out1cluster7) (1)

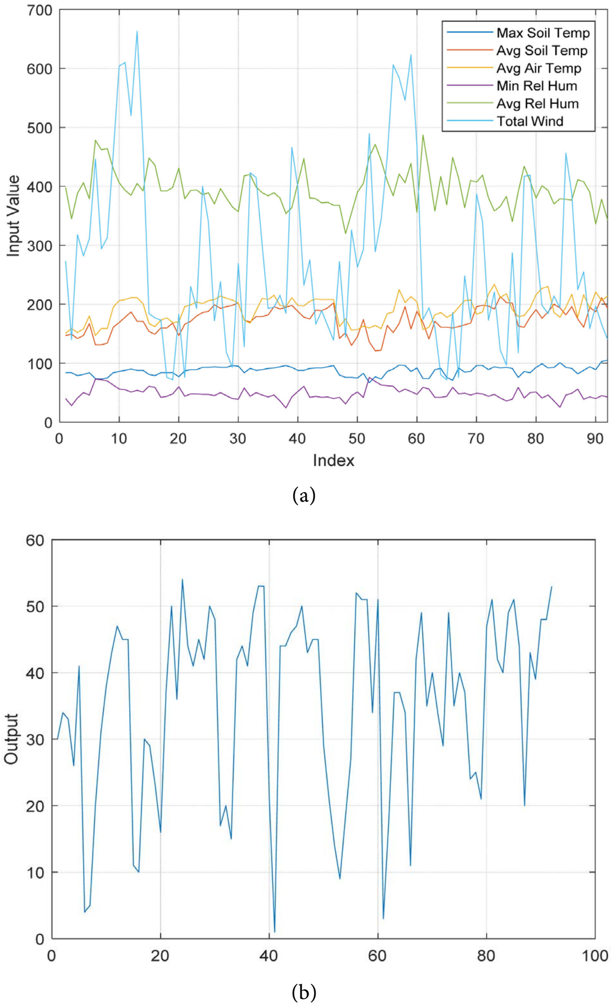
Figure 4. Few fuzzy rules.

4. Results

The profile of input and output variables against the index is shown in Figure 5(a) and Figure 5(b) respectively. The graphs reveal that input and output are uncorrelated.

First of all we modeled the relationship between the input and the output variable by clustering the data, where the cluster centers are used as a basis to define a Fuzzy Inference System (FIS). Here clustering is done to concise the representation of relationships embedded in the data. We apply subtractive clustering, which will make the appropriate number of clusters and the cluster centers taking radius of 0.5 shown in Figure 6. We got 7 cluster center for each input-output scatterplot, for example 7 centers of Daily-Evap vs. Max-Soil-Temp plot are: (92, 44), (90, 47), (77, 16), (84, 30), (84, 21), (79, 33) and (73, 5). Next FIS is created using the "subtracting clustering" under Matlab 18.

Here, 7 cluster centers and their range of influences are used as the input of FIS. The surface plot, relating each two inputs and output of the FIS is given in Figure 7.



(a)

(b)

Figure 5. Variation of input and output variables. (a) Input data; (b) Output data.

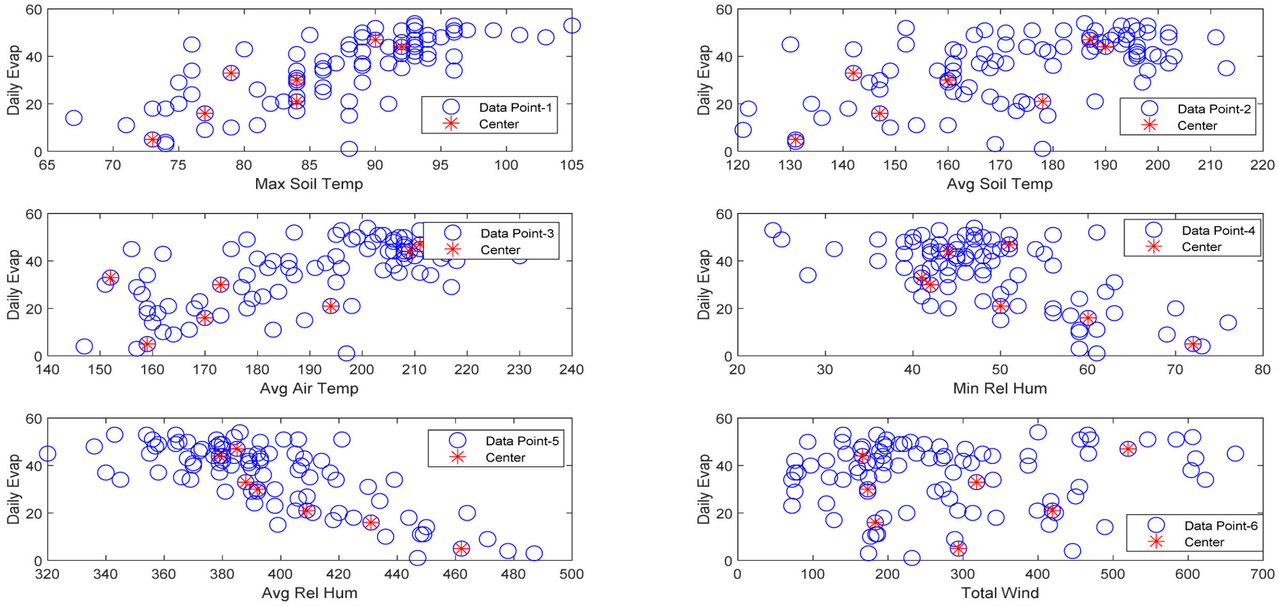


Figure 6. Scatterplot of input vs output dataset.

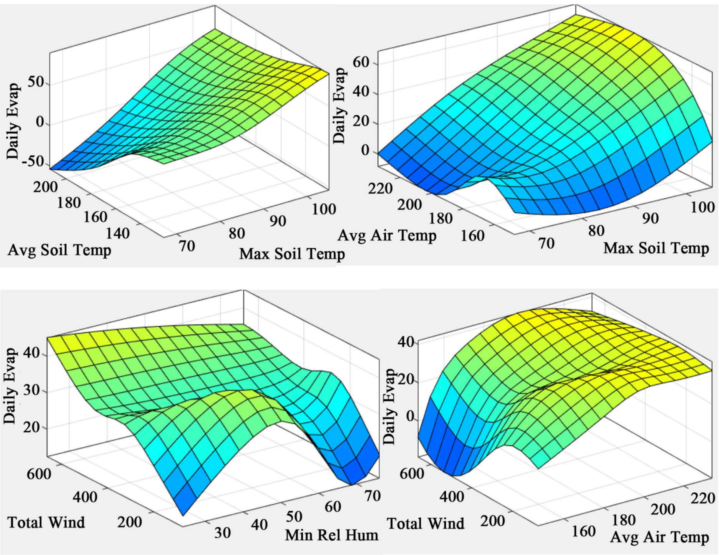


Figure 7. Input-output surface viewer.

The rule viewer of the FIS is shown in Figure 8, where each input has seven MFs, because of seven clusters of scatter-plot. Varying the magnitude of input data, the output of Figure 8 changes accordingly, where some numerical data relating input and output is shown in Table 2.

The output result of the paper is compared with ANN prediction, but the result of the paper provides more closed value to the target value. The comparative results are shown graphically in Figure 9 and the percentage of error between "target and our prediction" and "target and ANN prediction" is shown in Figure 10. Both Figure 9 and Figure 10 prove the superiority of our prediction model compared to ANN.

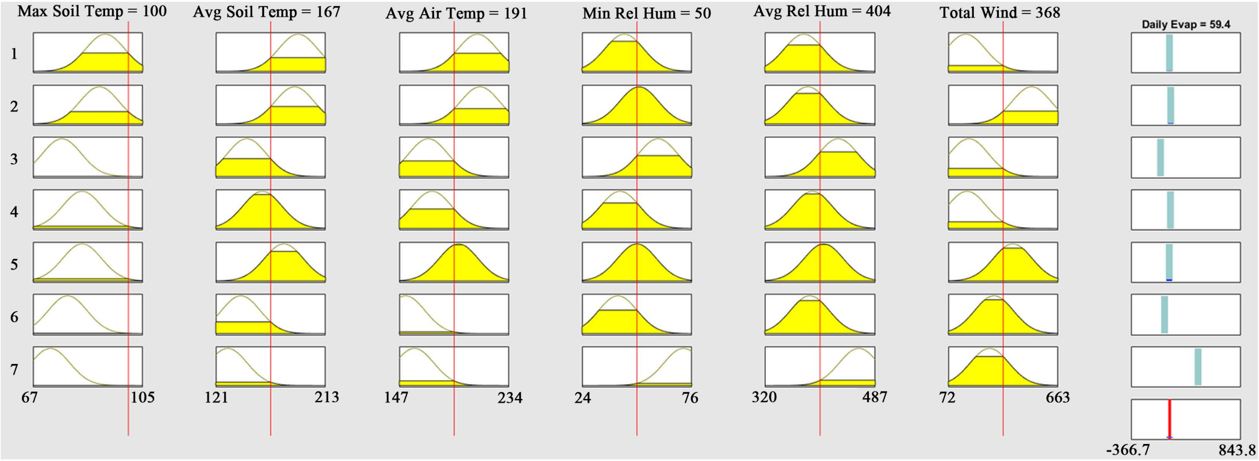


Figure 8. Rule viewer that simulates the entire FIS.

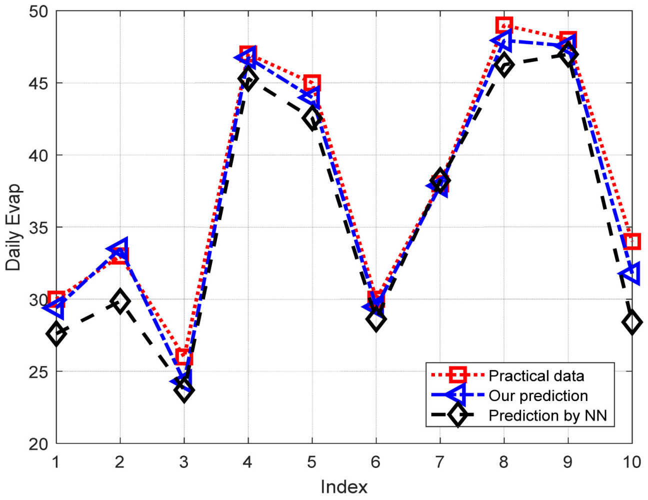


Figure 9. Comparison of among target value, NN prediction and our prediction.

Table 2. Daily evaporation prediction.

Max Soil Temp	Avg Soil Temp	Avg Air Temp	Min Rel Hum	Avg Rel Hum	Total wind	Daily Evap (Target value)	Neural Works prediction	Our Model Prediction
84	147	151	40	398	273	30	27.60808563	29.4
79	142	152	41	388	318	33	29.86248589	33.5
81	147	158	50	406	282	26	23.70377922	24.3
90	187	211	51	385	520	47	45.29627228	46.75
88	171	201	51	392	467	45	42.55496216	43.98
84	160	173	42	392	173	30	28.61962318	29.47
86	169	206	56	406	604	38	38.23920441	37.87
101	194	178	25	379	194	49	46.25722885	47.94
103	211	206	45	378	166	48	46.9826355	47.56
76	161	178	44	369	72	34	28.40890312	31.78

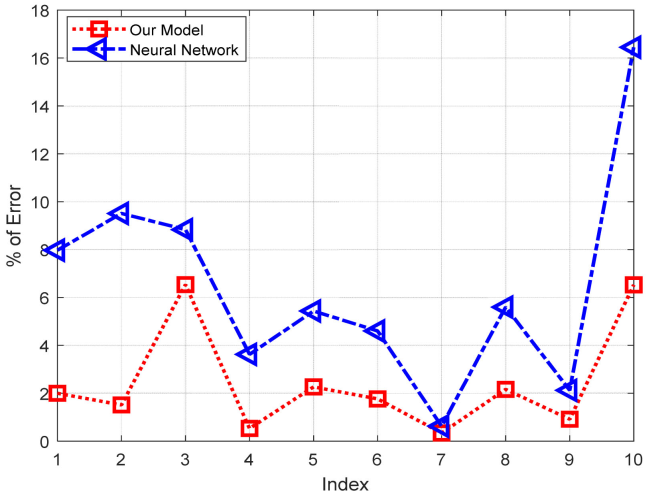


Figure 10. Comparison of percentage error.

5. Conclusion

In this paper, we take practical "evaporation of soil" as the target output relating with six environmental parameters. Our result shows better accuracy compared to previous work of ANN. In future we will combine FIS with ANN to acquire more accuracy compared to ANN. We still have the scope of using SVM, the K-means clustering algorithm, Fuzzy c-mean clustering, Naïve Bayes classifier, CART, etc. with FIS for comparison in context of accuracy and process time.

Conflicts of Interest

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

References

[1] Mirbagheri, M. (2010) Fuzzy-Logic and Neural Network Fuzzy Forecasting of Iran GDP Growth. *African Journal of Business Management*, **4**, 925-929.

[2] Liao, K.J., Li, X.F., Mu, C.X. and Wang, D. (2018) Short-Term Photovoltaic Power Prediction Based on T-S Fuzzy Neural Network. 2018 33*rd Youth Academic Annual Conference of Chinese Association of Automation* (*YAC*), Nanjing, 18-20 May 2018, 620-624. https://doi.org/10.1109/YAC.2018.8406448

[3] Ai, C., Jia, L.J., Hong, M. and Zhang, C. (2020) Short-Term Road Speed Forecasting Based on Hybrid RBF Neural Network with the Aid of Fuzzy System-Based Techniques in Urban Traffic Flow. *IEEE Access*, **8**, 69461-69470. https://doi.org/10.1109/ACCESS.2020.2986278

[4] Xie, Z.X., Wang, R.G., Wu, Z.H. and Liu, T. (2019) Short-Term Power Load Forecasting Model Based on Fuzzy Neural Network using Improved Decision Tree. 2019 *IEEE Sustainable Power and Energy Conference* (*iSPEC*), Beijing, 21-23 November 2019, 482-486.

[5] Akkurt, S., Tayfur, G. and Can, S. (2004) Fuzzy Logic Model for the Prediction of Cement Compressive Strength. *Cement and Concrete Research*, **34**, 1429-1433. https://doi.org/10.1016/j.cemconres.2004.01.020

[6] Abboud, N., Daher, A., Darwich, M., Nachar, S. and Kamali, W. (2019) Development of a New Real Time Epilepsy Prediction Approach Based on Adaptive Neuro Fuzzy Inference System. 2019 *Fifth International Conference on Advances in Biomedical Engineering* (*ICABME*), Tripoli, 17-19 October 2019, 1-4. https://doi.org/10.1109/ICABME47164.2019.8940305

[7] Xu, B.H., Sun, Y.L., Wang, H.Y. and Yi, S.M. (2019) Short-Term Electricity Consumption Forecasting Method for Residential Users Based on Cluster Classification and Backpropagation Neural Network. 2019 11*th International Conference on Intelligent Human-Machine Systems and Cybernetics* (*IHMSC*), 24-25 August 2019, Hangzhou, 55-59.

[8] Sanjaya, C., Liana, M. and Widodo, A. (2010) Revenue Prediction Using Artificial Neural Network. 2010 *Second International Conference on Advances in Computing, Control, and Telecommunication Technologies*, Jakarta, 2-3 December 2010, 97-99. https://doi.org/10.1109/ACT.2010.53

[9] Pahwa, K. and Agarwal, N. (2019) Stock Market Analysis Using Supervised Machine Learning. 2019 *International Conference on Machine Learning*, *Big Data*, *Cloud and Parallel Computing* (*Com-IT-Con*), Faridabad, 14-16 February 2019, 197-200. https://doi.org/10.1109/COMITCon.2019.8862225

[10] Li, Y.L., Li, B.B. and Yin, C.Y. (2010) Modulation Classification of MQAM Signals Using Particle Swarm Optimization and Subtractive Clustering. *IEEE* 10*th International Conference on Signal Processing Proceedings*, Beijing, 24-28 October 2010, 1537-1540. https://doi.org/10.1109/ICOSP.2010.5656376

[11] Gu, L. and Lu, X.L. (2012) Semi-Supervised Subtractive Clustering by Seeding. 2012 9*th International Conference on Fuzzy Systems and Knowledge Discovery* (*FSKD* 2012), Sichuan, 29-31 May 2012, 738-741. https://doi.org/10.1109/FSKD.2012.6234240

[12] Barchinezhad, S., Eftekhari, M. and Sanatnama, H. (2013) A New Feature Ranking Criterion Based on Density Function of Subtractive Clustering. 13*th Iranian Conference on Fuzzy Systems* (*IFSC*), Qazvin, 27-29 August 2013, 1-4. https://doi.org/10.1109/IFSC.2013.6675624

[13] Gu, L. (2016) A Novel Subtractive Clustering by Using K-Harmonic Means Clustering for Initialization. 7*th IEEE International Conference on Software Engineering and Service Science* (*ICSESS*), Beijing, 26-28 August 2016, 840-843. https://doi.org/10.1109/ICSESS.2016.7883197