Journal of Engineering and Applied Sciences 14 (12): 4035-4042, 2019
ISSN: 1816-949X

Fuzzy Inference System Model from Non-Fuzzy Clustering Output

Nur Atiqah Binti Hamzah and Sie Long Kek
Department of Mathematics and Statistics, University Tun Hussein Onn Malaysia,
Batu Pahat, Malaysia

Abstract: Fuzzy Inference System (FIS) is a process of mapping input into the desired output using fuzzy logic theory where decisions can be made or patterns are discerned. This study aims to discuss on how non-fuzzy clustering output can be used to construct a model of FIS. Here, the proposed idea is to show the efficient use of the FIS as a prediction model for the data classification. In this study, employment income, self-employment income, property and transfer received are taken into account for clustering the household income data. Then, the FIS prediction model is built using the center values of clusters formed and the output of FIS is compared to the original cluster in which the best fit prediction model to the data is determined. In conclusion, the best prediction model in identifying income class is discovered based on the Root Mean Square Error (RMSE) value computed.

Key words: Household income data, fuzzy inference system, k-means clustering, root mean square, prediction model, Root Mean Square Error (RMSE)

INTRODUCTION

Clustering is one of well-known methods in grouping the set of data. Cluster analysis is a formal study of methods on using algorithm to group the objects with similar characteristics into the same cluster. This solution method is sometimes known as unsupervised classification where there is no information about the class labels or the number of clusters known. The goal of clustering is to minimize the distance within the same group and to maximize the distance from the other clusters. Clustering has been used as a stand-alone tool to give the insight of data distribution and also as preprocessing step for other algorithms. There are many applications of clustering in the real world problems such as data documentations, image processing, pattern recognition and spatial data analysis.

A quality of clustering depends on similarity measure used and its implementation. Besides, the quality of clustering also depends on its ability to discover the hidden patterns from the data. There are many approaches in clustering techniques such as partitioning algorithms, hierarchy algorithms, density-based and model-based algorithms. These approaches lead to different result in formation of clusters. In literature, k-means clustering which was introduced by MacQueen (1967) is the popular techniques for doing clustering according to the cluster mean (MacQueen, 1967).

On the other hand, Fuzzy Inference System (FIS) which is the process of formulating the mapping from a given input to a desired output by using fuzzy logic concept (Zadeh, 1965) is applied in which decisions can be made or patterns are discerned. Rules form is the basis of the fuzzy logic which comprises the rule-based system originated from sources other than that of human experts. The system uses If-Then rule-based system which is If antecedent, then consequent.

Where the antecedents express an inference or the inequality and the consequents are those that can be inferred as an output if the antecedent inequality is satisfied. Note that the number of rules is increased exponentially with the dimension of the input space.

Furthermore, a fuzzification interface, a rule-base, a database, a decision-making unit and a defuzzification interface are crucial components in FIS (Acqua and Abbondanti, 2003). Mamdani and Takagi-Sugeno systems are 2 examples of FIS that are widely used in various fields such as automatic control, data classification, decision analysis, expert systems and computer vision. Mamdani's method is among the first control systems built by using fuzzy set theory. Meanwhile Sugeno-type system can be used to model any inference system in which the output of membership functions is either linear or constant function (Kaur and Kaur, 2012). The Takagi-Sugeno Model was proposed by Takagi, Sugeno

Corresponding Author: Nur Atiqah Binti Hamzah, Department of Mathematics and Statistics, University Tun Hussein Onn Malaysia, Batu Pahat, Malaysia

and Kang which is known as an effort to formalize a system approach to generate fuzzy rules from an input-output data set.

FIS can have either fuzzy input or crisp input and the output is always in fuzzy sets. In data mining, fuzzy methods are used in a certain process such as database queries, data preparations, modelling phase and evaluation phase. Fuzzy system is also known as an intelligent system with a potential to produce models which are more comprehensive, simplicity and more robust. Hence, fuzzy methods are useful for pre-and post-processing of data (Devi and Rani, 2013). For instant, there were research used clusters before membership functions were constructed (Al-Shammaa and Abbod, 2014; Kalpana and Senthil Kumar, 2011).

In this study, k-means clustering is used to identify the number of groups of household income data. By using the distance measures which are square Euclidean distance, cityblock, correlation and cosine, the best clustering result is determined in order to design fuzzy rules and fuzzy membership functions. Three shapes of membership functions used in this study are triangle, trapezoidal and Gaussian membership functions. Moreover, the class of the household income in the FIS prediction model built could be identified when the input value is added into the system. The results are then compared with the output value from clustering process.

Problem statement: Household incomes data which can be used to measure poverty or economy level of the country are really important for each country. Previously, there is no other studies in Malaysia which is focusing on finding cluster based on household income data. Due on the number of groups based on household income data collected is unknown and on how each household income data can be put into one group is also unclear, the decision making on the policy regarding to the national development would not be satisfied. Therefore, a comprehensive analysis on the household income data is necessary required, not only for the residents but also for the government to realize the level of expenditure and saving for the future.

Since, the number of data collected on household income is large, a suitable technique of data mining should be applied for data grouping. However, different initial points of starting can give in different final clusters which makes the undesired results instead. In addition, the program or software used usually ought to be rerun with the same value of the number of clusters or different value of the number of clusters. This is simply to compare the results and to select the best results.

Therefore, for a meaningful result on clustering output, an intelligent tool or a technique should be applied. Here, FIS approach is implemented on non-fuzzy clustering output, so as the best model for the household income data can be identified.

MATERIALS AND METHODS

In order to achieve the objective of this study, the methods of k-means clustering process and FIS using MATLAB Software are further discussed in this section.

k-means clustering: In k-means clustering, the centroids are selected randomly for each cluster. The distance of data points to the centroid is calculated using the selected distance measure. Each point is placed in the group based on the shortest distance to the centroid. The centroid value is then updated by finding the means of the elements in the groups. The formal k-means clustering algorithm 1 is given as follows:

Algorithm 1; Clustering algorithm:

Step 1	Choose a value of the number of clusters k
Step 2	Select points as initial centroids
Step 3	Calculate the distance of points to centroids
Step 4	Assign points to the group based on distance computed
Step 5	Update the centroid by finding the means
Step 6	Repeat steps 3-5
Step 7	Stop the process when there is no more changes in each group

Figure 1 shows the k-means clustering process involved in this study. The process of clustering with the different types of distance measure is applied to the k-means clustering algorithm.

In this study, all computation processes are carried out using MATLAB R2015a Software. Distance between data points and the center is calculated using distance formula as shown below: squared Euclidean distance is defined by Bora and Gupta:

$$d(x, y) = (x_n - x_c)^2 + (y_n - y_c)^2 \quad (1)$$

Where:

x_n = Value of variable x for n observation
y_n = Value of variable y for n observation
x_n = Center value for variable x
y_n = Center value for variable y

Manhattan or cityblock distance is defined by Dalfo *et al.* (2007):

$$d(x_n, x_c) = \sum |(x_n - x_c)| \quad (2)$$

Where:

x_n = Value for variable x for n observation
x_c = Center value for variable x

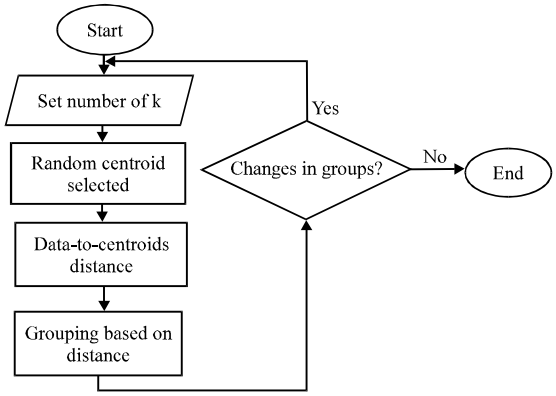

Fig. 1: k-means clustering process

Cosine formula is defined by Huang (2008):

$$d(x_n, x_c) = 1-\frac{(x_n x'_c)}{(\sqrt{x_n x'_n})(\sqrt{x_c x'_c})} \quad (3)$$

Where:

x_n = Value for variable x for n observation

x_c = Center value for variable x

Correlation distance is defined by Taghva and Veni:

$$d(x_n, x_c) = 1-\frac{(x_n-\bar{\bar{x}}_n)(x_c-\bar{\bar{x}}_c)}{\sqrt{(x_n-\bar{\bar{x}}_n)(x_n-\bar{\bar{x}}_n)}\sqrt{(x_c-\bar{\bar{x}}_c)(x_c-\bar{\bar{x}}_c)}} \quad (4)$$

Where:

x_n = Value for variable x for n observation

x_c = Center value for variable x

After the distance for each data point is calculated, the data are grouped into respective cluster based on the shortest distance. Then, the new center is updated based on the means of each cluster variables. This iteration process is continued until there is no changes in groups which show the stable cluster solution is achieved. The same distance measure is used until the end of the clustering process. In order to get the output by using other distance measures, the clustering process would be started from the beginning. Total sum of distance is computed and the smallest value indicates the best cluster as the differences between data points n to the group center is small.

Besides that, Silhouette index and cluster figure are used to compare the performance such that the number of clusters k gives the best solution can be determined. Silhouette coefficient quantifies the quality of clustering as calculated from:

$$s(x) = \frac{b(x)-a(x)}{\max\{a(x), b(x)\}} \quad (5)$$

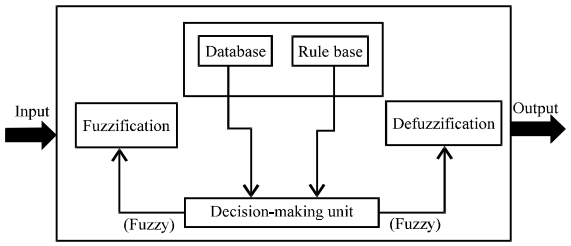

Fig. 2: Building fuzzy inference system steps

Table 1: Input variables and fuzzy numbers

Input variables	Class of income
Employment income	Low, medium, high
Self-employment income	Low, medium, high
Property	Low, medium, high
Transfer received	Low, medium, high

where, a(x) is the average distance of x to all other vectors in the same clusters while b(x) is the minimum average distance of x to the vectors in other clusters (Izenman, 2009). The Silhouette index has a range [-1, 1]. The value of -1 shows the poorly matched cluster and the value of 1 shows the well-matched to the cluster. If many points have a high value of Silhouette index, then the clustering solution is appropriate. The Silhouette clustering evaluation criterion can be used with any distance metric.

Meanwhile the cluster figure is based on the Silhouette value of each observation n. The cluster figure can be used to compare the effectiveness for all the distance measure in performing k-means clustering to the dataset (Izenman, 2009).

Building fuzzy inference system: After the cluster solutions are obtained, membership function and fuzzy rules will be designed. The number of rules for FIS that will be designed according to fuzzy operator is based on the clusters formed. The inputs for FIS are the variables that have been clustered and the output is the class of the household incomes. Figure 2 shows the steps in constructing a FIS in MATLAB Software with fuzzy toolbox.

The fuzzy verdict mechanism separately infers the possibility of the household income for each instance in fuzzification and transfers the possibility into the form of sentences. There are four variables, namely, employment income, self-employment income, property and transfer received are chosen as the input variables. The parameters of the class of income are listed in Table 1.

Algorithm 2; The algorithm of fuzzy verdict mechanism is given as follows:

INPUT

Enter the fuzzy set for employment income, self-employment income, property and transfer received.
OUTPUT
Class of household income = {low, medium, high}
METHOD
Begin
Step 1: Enter the crisp values of all inputs
Step 2: Set the membership function
Step 3: Built fuzzy number for input and output
Step 4: Fuzzy inference are executed by Mamdani's or Takagi-Sugeno's methods
Step 5: Centroid method is applied by

$$z^* = \int \frac{\mu_{\varsigma}(z) z dZ}{\mu_{\varsigma}(z) dZ}$$

Step 6: Present the output in human nature language
End

The data are tested using the FIS that had been built. The Root Mean Square Error (RMSE) will be used to compare the output from FIS with the k-means clustering output. The RMSE is defined by Chai and Draxler, (2014):

$$RMSE = \sqrt{\frac{\Sigma(e)^2}{n}} \quad (6)$$

where, e is error value obtained by finding differences between output from FIS with output from k-means clustering.

RESULTS AND DISCUSSION

The household income data with n = 13,233 observations were used in this study to achieve the research objectives. Only 4 attributes were selected which are employment incomes, self-employment incomes, property income and transfers received as shown in Table 2.

Figure 3 shows the household income data distributed and there is no clearly number of clusters in the data. Therefore, to perform k-means clustering on the data, the number of clusters k need to be set. The clustering process was carried out with k = 2 and k = 3 using 4 distance measures which are square Euclidean distance, cityblock, correlation and cosine. Different distance measures were used in order to select the best cluster output of the household incomes. These results are shown in Table 3.

Here, due on the smallest total sum of distance, the best clustering output with the number of clusters k = 3 by using cosine distance has been chosen to build FIS. In addition, strong mean Silhouette index value which is closer to +1 shows that the cluster formed is good. The cluster figure is also a guideline on how well the data was placed in each group; either it was correctly matched or not. Later on, fuzzy rules and membership function would be set based on the selected clustering output and the collected center value of each cluster.

By using the MATLAB fuzzy toolbox, 3 shapes of membership functions, namely, triangular, trapezoidal and Gaussian shall be determined. Then, these shapes of membership functions were applied to 2 types of FIS which are Mamdani and Takagi-Sugeno. Here, the FIS rules base were set according to the cluster center from k-means clustering output which is shown in Table 4.

Based on the clustering output, all variables have three classes which are low, medium and high. Let us introduce the following notations before building the fuzzy rule:

Table 2: Household income data information

Name of dataset	Household income data
Dataset characteristics	Multivariate
Number of instances	13,233
Number of attributes	4
Attribute characteristics	Real
Area	Economy

Table 3: k-means clustering output

Value of cluster k	Distance measure	Total sum of distance	Mean Silhouette index
2	Square Euclidean distance	2.80465×10^{13}	0.7442
2	Cityblock	4.48019×10^{8}	0.3888
2	Correlation	1606.1	0.8023
2	Cosine	981.352	0.7928
3	Square Euclidean distance	2.22862×10^{13}	0.7733
3	Cityblock	3.66337×10^{8}	0.3581
3	Correlation	677.538	0.8638
3	Cosine	448.486	0.8489

Table 4: Center value for each clusters

Variable (V)	Cluster 1	Cluster 2	Cluster 3
1	0.099614	0.962584	0.128909
2	0.349967	0.138661	0.939063
3	0.030969	0.026469	0.020149
4	0.865700	0.03894	0.124054

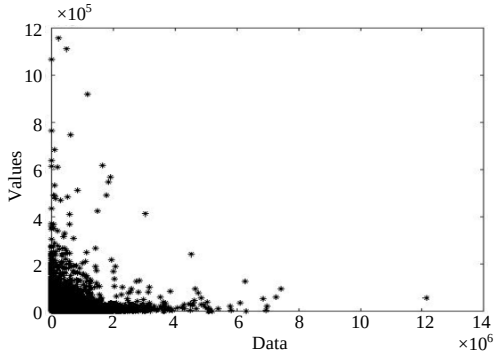

Fig. 3: Data distribution

Table 5: Fuzzy inference system model

Models	Type of FIS	Input membership function	Output membership function	And	Or	Aggregation	Defuzzification
1	Mamdani	Triangle	Triangle	min	max	max	Centroid
2	Mamdani	Trapezoidal	Triangle	min	max	max	Centroid
3	Mamdani	Gaussian	Triangle	min	max	max	Centroid
4	Takagi-Sugeno	Triangle	Constant	min	max	-	Weighted average
5	Takagi-Sugeno	Trapezoidal	Constant	min	max	-	Weighted average
6	Takagi-Sugeno	Gaussian	Constant	min	max	-	Weighted average

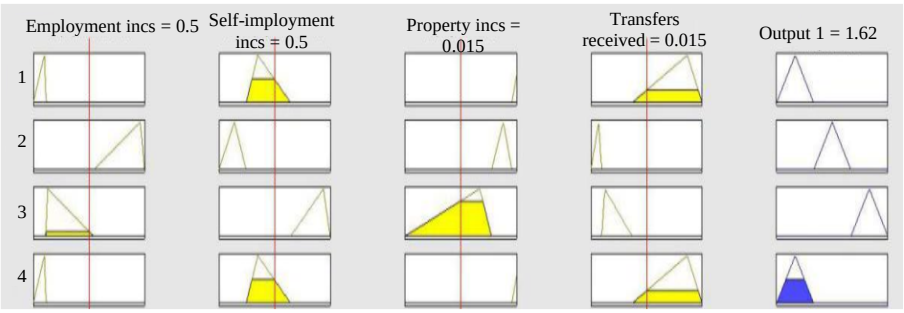

Fig. 4: Rules view for Model 1

- V_1 is employment income
- V_2 is self-employment income
- V_3 is property income
- V_4 is current transfer received
- C_1 is class 1 with value of output less than or equal to 1.5
- C_2 is class 2 with value of output more than 1.5-2.5
- C_3 is class 3 with value of output >2.5

Hence, the rules base of the FIS are designed as follows:

Rule 1: If V_1 is low and V_2 is medium and V_3 is high and V_4 is high then household income is in C_1.

Rule 2: If V_1 is high and V_2 is low and V_3 is medium and V_4 is low then household income is in C_2.

Rule 3: If V_1 is medium and V_2 is high and V_3 is low and V_4 is medium then household income is in C_3.

Rule 4: If V_1 is low or V_2 is medium or V_3 is high or V_4 is high, then household income is in C_1.

Rule 5: If V_1 is high or V_2 is low or V_3 is medium or V_4 is low then household income is in C_2.

Rule 6: If V_1 is medium or V_2 is high or V_3 is low or V_4 is medium then household income is in C_3.

Note that if the variable observation has the same characteristics with the rules listed it can be classified into the selected class. Otherwise, the fuzzy system will help to define which class does the variable observation belongs to. The income classes which are low, medium and high were put into fuzzy expression based on the shape of membership function where the range of each class of the selected variable observation was determined.

Table 5 shows the result obtained by using the MATLAB fuzzy toolbox. There are 6 models which cover input membership function, output membership function, the and operator, the or operator, aggregation and defuzzification process. Among these 6 models used, three models were Mamdani FIS and the other 3 models were Takagi-Sugeno FIS. The outputs were further analyzed to investigate how these models were differ to each other.

Figure 4 the rules view for the model that had been designed. It shows the membership functions of inputs and the value of outputs after defuzzification process. The output depends the value on the shape of membership function, fuzzification, aggregation and defuzzification of the models.

For the accuracy of the output obtained from FIS, the output would be compared with the original output that is obtained from k-means clustering process. To do so, the original data were transformed into cosine value. It is because the best clustering output was obtained using cosine distance. The transformed data were put into array form in MATLAB. The data were tested for each model using the function evalfis as shown in algorithm 3. To evaluate the efficiency of the output obtained, the RMSE function is built in the m-file as shown in algorithm 4 (Table 6). The RMSE values were used to find the error

Algorithm 3; Commands to evaluate the class of the data:

```
Ismatl = readfis ('FISTSKqauss')
%k =
Out 1 = evalfis ([1, 0.913545457648601, ...,], fismat 1
```

Algorithm 4; Commands for RMSE function in m-file:

```
Function rmse (data, estimate)
R = scrt (sum((data (:)-estimate (:)).  2/nume 1 (data)))
```

Algorithm 5; Commands for RMSE evaluation:

```
>>  rmse (Oldoutput, newoutput)
R =
```

Table 6: RMSE value for six different models

Models	Fuzzy inference system	Membership function	RMSE
1	Mamdani	Triangle	1.4008
2	Mamdani	Trapezoidal	1.4910
3	Mamdani	Gaussian	1.1825
4	Takagi-Sugeno	Triangle	1.4265
5	Takagi-Sugeno	Trapezoidal	1.4235
6	Takagi-Sugeno	Gaussian	0.9788

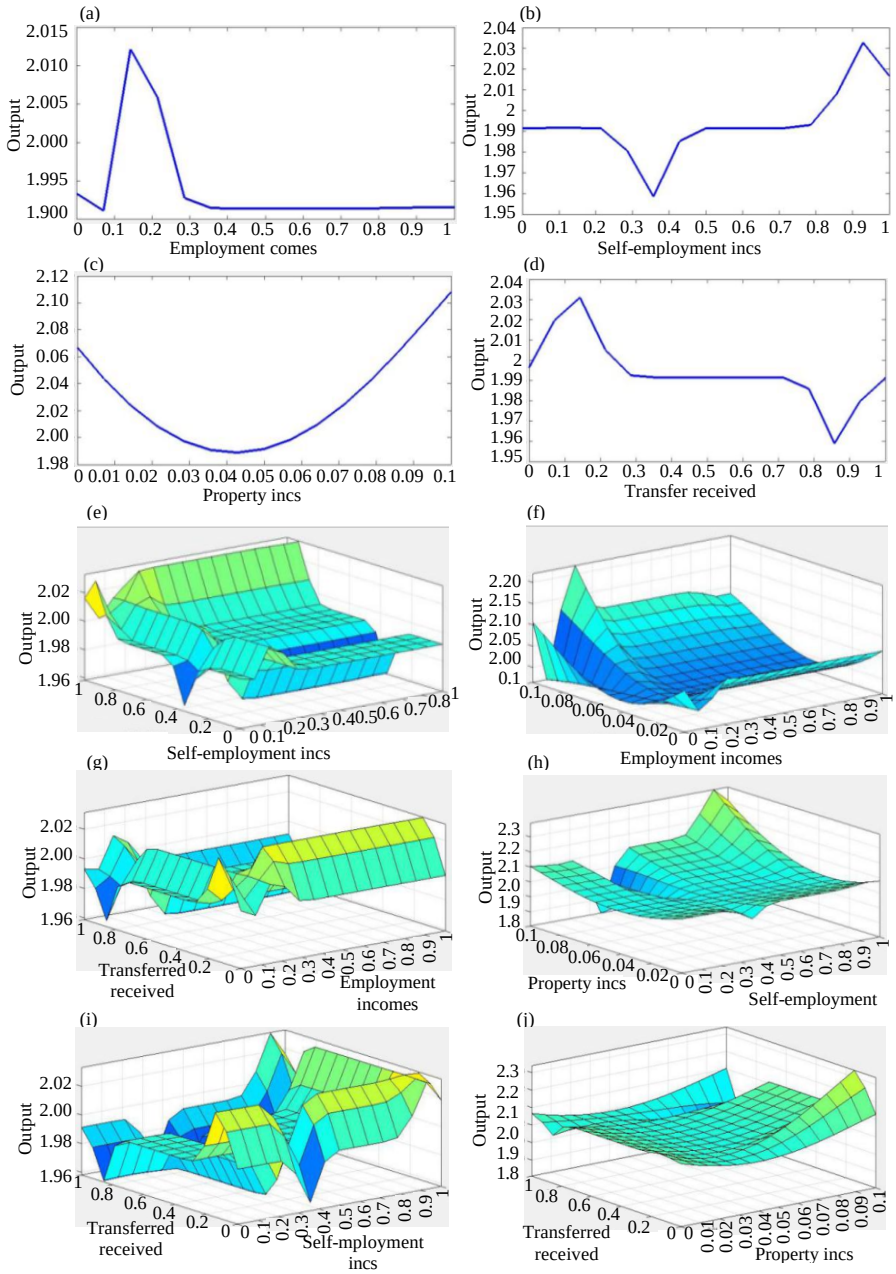

Fig. 5: Input-output relationship graphic view: a)Employment income versus output value; b)Self-employment income versus output value; c)Property versus output value; d) Transfer received versus output value; e)Input: Employment and self-employment; f)Input: Employment income and property income; g) Input: Employment income and transfer received h)Input: Self-employment and property income; I) Input: Self-employment and transfer received and j) Input: Property income and transfer received

between the output value using FIS Model and original cluster. Then, the RMSE of this error was calculated using the command as shown in algorithm 5. Algorithm 5, R = 1.4008 is an example, of the RMSE value obtained by comparing the new outputs from the FIS model with the original output from k-means clustering. Since, the original clusters were defined to have range from 1-3. When the data were put into the FIS that had been designed, the outputs were in 3 classes of household incomes groups which are low, medium and high. Hence, the smallest error shows the efficient prediction model. The result of RMSE is shown in Table 6.

Notice that Model 3 has the smallest RMSE value which is 1.1825 for Mamdani FIS. Meanwhile, Model 6 has the smallest RMSE value for Takagi-Sugeno FIS which is 0.9788. Both models used Gaussian membership function. However, RMSE value for Model 6 is smaller than Model 3. This indicates that model 6 gives the best prediction of household income class and the output from the model did not have big difference from the original output. When the membership function used was triangle, the RMSE value for Mamdani FIS was smaller than Takagi-Sugeno FIS. However, when the membership function used was trapezoidal, the RMSE value for Takagi-Sugeno is better than Mamdani FIS.

Therefore, FIS is suitable to be used in determining which class for the data belong to because k-means clustering will always give different result due to its initial point. Thus, FIS approach is to make the cluster output more useful and easy to understand. The selection of the center based on one cluster process which then has been used in FIS was to make sure the process to grouping data was standardized and not changing or influenced by calculation process during clustering. The relationship between the input to the output can be shown graphically in the surface view in Fig. 8.

CONCLUSION

This study was an interesting topic because number of groups of household incomes were unknown. It means that there is no clear information how a person can be categorized as low, middle or high group of income. In this study, a normal k-means clustering was used as its simplicity of computation. k-means clustering on household incomes data was managed to find how many groups of incomes that exist among the Malaysian community.

The FIS for household income was successfully built by using k-means clustering output. It shows that k-means clustering can be a beneficial platform to draw a good idea before building FIS. Variation did exist in every model as the shapes of membership function used were different. However, it can be concluded that Gaussian membership function gives the best output for both Mamdani and Takagi-Sugeno models. The least value in RMSE formed was considered to select the best model. It shows that the classification done by using FIS was still obeying the clustering concept that has been carried out on the first place.

RECOMMENDATIONS

In future research it is recommended if new technique of clustering can be applied to the household income data. Thus, the effectiveness of different techniques can be compared. Besides that by finding the prediction model of household income class was actually a basic idea to expand the model to be a system and be implemented in web system. It is crucial to identify in which class of economy someone gains to increase their effort to improve their life.

Besides that, these research techniques should be applied to other data too. It is important to know if the technique fit with other data or not. FIS and other fuzzy logic approach should be applied as a tool for classification and other applications. This technique can be widely used in other field as it is compatible with human logic.

ACKNOWLEDGEMENT

Researchers would like to express the greatest gratitude to the Office of Research innovation, Commercialization and Consultancy (ORICC), University Tun Hussein Onn Malaysia (UTHM) for providing the financial support in carrying out this study through the Postgraduate Research Grant (GPPS) VOT 606. Also, researcher would extremely thankful to reviewers for their valuable remarks.

REFERENCES

Acqua, G.D. and R.L.F. Abbondanti, 2003. Adaptive neuro fuzzy inference system for highway accidents analysis. J. Fuzzy Syst., 1: 1-6.

Al-Shammaa, M. and M.F. Abbod, 2014. Automatic generation of fuzzy classification rules from data. Proceedings of the International Conference on Neural Networks-Fuzzy Systems (NN-FS'14), March 15-17, 2014, Venice, Italy, pp: 74-80.

Bora, D.J. and A.K. Gupta, 2014. A comparative study between fuzzy clustering algorithm and hard clustering algorithm. Int. J. Comput. Trends Technol., 10: 108-113.

Chai, T. and R.R. Draxler, 2014. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)-Arguments against avoiding RMSE in the literature. Geoscient. Model Dev., 7: 1247-1250.

Dalfo, C., F. Comellas and M.A. Fiol, 2007. The multidimensional manhattan network. Electron. Notes Discrete Math., 29: 383-387.

Devi, M.K. and M.U. Rani, 2013. Fuzzy inference system and its application. Intl. J. Eng. Sci. Res., 4: 1248-1250.

Huang, A., 2008. Similarity measures for text document clustering. Proceedings of the Sixth New Zealand Conference on Computer Science Research Student (NZCSRSC2008), April 14-18, 2008, University of Canterbury, Christchurch, New Zealand, pp: 49-56.

Izenman, A.J., 2009. Modern Multivariate Statistical Techniques: (Regression, Classification and Manifold Learning). Springer, Berlin, Germany, ISBN:978-0-387-78188-4, Pages: 731.

Kalpana, M. and A.V. Senthil Kumar, 2011. Fuzzy Expert System for Diabetes using Fuzzy Verdict Mechanism. Int. J. Advanced Net. Applic., 3: 1128-1134.

Kaur, A. and A. Kaur, 2012. Comparison of mamdani-type and sugeno-type fuzzy inference systems for air conditioning system. Intl. J. Sof. Comput. Eng., 2: 323-325.

MacQueen, J., 1967. Some methods for classification and analysis of multivariate observations. Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability, Volume 1, January 17-20, 1967, California, USA., pp: 281-297.

Zadeh, L.A., 1965. Fuzzy sets. Inform. Control, 8: 338-353.