MALAYSIA HOUSEHOLD INCOMES CLASSIFICATION PREDICTION WITH *K*-MEANS CLUSTERING AND FUZZY INFERENCE SYSTEM

NUR ATIQAH BINTI HAMZAH

A thesis submitted in fulfillment of the
requirement for the award of the degree of
Master of Science

Faculty of Applied Science and Technology
Universiti Tun Hussein Onn Malaysia

AUGUST 2018

Every challenging work needs self-efforts as well as guidance.

For my beloved mother and father, thank you for your moral supports during my hard time. Your affections, encouragement, prays of day and night are the secrets of each success and honor I achieved.

For my respected supervisor, Dr.Kek Sie Long, thank you for being my superb mentor.

For my family and friends, thank you for always be there for me.

This thesis is for all of you.

ACKNOWLEDGEMENT

I am very grateful to the Most Merciful Allah SWT for His countless gifts to me.

It is my privilege to thank my family especially to my parents for their constant encouragement in completing this study either in mentally or spiritual aspects. I always appreciate your supportive words.

I want to express my deepest thanks to my supervisor Dr. Kek Sie Long for his guidance. It is a great honour to be his Master's Degree student.

My sincere gratitude to the Universiti Tun Hussein Onn Malaysia (UTHM) for letting me of being a student here and allowing me to do the research using GPPS sponsorship.

I am extremely thankful to all my friends who always being my good supporters.

Nur Atiqah Hamzah

August 2018

ABSTRACT

The economy level of the citizen has become a main concern for Malaysia as a developing country to improve the living status. On this point of view, the household income data would be a very useful information to measure the economic status of the population in Malaysia. This study aims to build a classification prediction of household incomes using fuzzy inference system (FIS) from the *K*-means clustering outputs. Thus, this study focuses on three main objectives which are (a) To apply *K*-means clustering on household incomes data, (b) To propose the prediction of household incomes classification using FIS, and (c) To analyze and validate the classification solution for household incomes and to compare with discriminant analysis. Initially, the number of groups in the household income data is determined by using *K*-means clustering. Accordingly, the outputs from *K*-means clustering are used to identify the membership functions, namely, triangle, trapezoidal and Gaussian membership functions. Furthermore, FIS models for each membership function are built for the household income class prediction based on clustering outputs. For verification, the root mean square error (RMSE) value for each FIS model is calculated and the percentage of data correctly classified using the FIS models built is compared with the discriminant analysis output. As a result, it is found that Mamdani FIS model with Gaussian membership function is the best model with the RMSE is 1.0396, while the percentage of data correctly classified is 64.9989%. In conclusion, the classification prediction of household incomes discussed in this thesis could identify the predicted class of household income in a tractable way and the efficiency of the technique used in this thesis for classification prediction of household income is highly recommended.

ABSTRAK

Taraf ekonomi masyarakat telah menjadi perhatian utama bagi Malaysia sebagai sebuah negara yang membangun untuk meningkatkan status hidup masyarakat. Dari sudut pandangan ini, data pendapatan isi rumah adalah menjadi maklumat yang sangat berguna untuk mengukur status ekonomi populasi di Malaysia. Kajian ini bertujuan untuk membina ramalan klasifikasi pendapatan isi rumah menggunakan sistem inferens kabur (FIS) dari hasil kajian pengklusteran *K*-min. Oleh itu, kajian ini menumpukan pada tiga objektif utama iaitu (a) Untuk menerapkan pengklusteran *K*-min ke atas data pendapatan isi rumah, (b) Untuk mencadangkan ramalan klasifikasi pendapatan isi rumah menggunakan FIS, dan (c) Untuk menganalisis dan mengesahkan penyelesaian klasifikasi untuk pendapatan isi rumah dan membandingkan penyelesaian tersebut dengan analisis diskriminasi. Pada peringkat permulaan kajian ini, jumlah kumpulan pendapatan isi rumah dikenalpasti menggunakan kaedah pengklusteran *K*-min. Selepas itu, hasil daripada proses pengklusteran *K*-min digunakan untuk mengenalpasti fungsi keahlian, iaitu fungsi keahlian segi tiga, trapezoid dan Gaussian. Seterusnya, model FIS bagi setiap fungsi keahlian dibina untuk ramalan kelas pendapatan isi rumah berasaskan daripada hasil pengklusteran. Bagi tujuan verifikasi, nilai ralat punca min kuasa dua (RMSE) bagi setiap model FIS dikira dan peratusan data yang diklasifikasikan dengan betul menggunakan model FIS dibandingkan dengan keputusan klasifikasi menggunakan analisis diskriminan. Daripada hasil kajian ini, didapati bahawa model Mamdani FIS dengan fungsi keahlian Gaussian adalah model terbaik dengan RMSE adalah 1.0396, manakala peratusan data yang dikelaskan dengan betul adalah 64.9989%. Kesimpulannya, ramalan klasifikasi pendapatan isi rumah yang dibincangkan dalam tesis ini dapat mengenalpasti kelas pendapatan isi rumah yang diramalkan dengan cara yang mudah dan kecekapan teknik yang digunakan dalam tesis ini untuk ramalan klasifikasi pendapatan isi rumah adalah sangat disyorkan.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PUBLICATIONS

The following papers (which have been published or accepted for publication) were completed during Master's Degree candidature.

Hamzah, N.A., Kek, S.L., and Saharan, S. (2017). The Performance of *K*-Means and *K*-Modes Clustering to Identify Cluster In Numerical Data, J*ournal of Science and Technology*, 9 (3), 25-32.

Hamzah, N.A, & Kek,S.L. (2017). A Comparative Analysis: Effect of Distance Measures to *K*-means Clustering Output. *Journal of Engineering and Applied Science.*

Hamzah, N.A, & Kek, S.L. (2018). Fuzzy Inference System Model from Non-Fuzzy Clustering Output. *Journal of Engineering and Applied Science*

CHAPTER 1

INTRODUCTION

1.1 Background of study

The household is defined as a group of people who live in the same family to share the same accommodations and meals. Household incomes are the combined incomes of all members of the household. The source of incomes might vary such as salaries and wages, retirement income, self-employment and owned property incomes (Varjonen and Aalto, 2005). For a better understanding, the household income survey was conducted by the Department of Statistics Malaysia (DOSM) since 1973. In 1987, the basic amenities survey was also conducted with a household income survey. There are five main objectives of the survey, which are (a) to measure the economic well-being of the population; (b) to collect information on income distribution pattern of household classified by various socio-economic characteristics; (c) to identify the poor groups; (d) to collect information on basic amenities of household; and (e) to study the effects of the implementation of national development program ("Household Income and Expenditure", 1987). Thus, identifying groups of household incomes in Malaysia using clustering technique and fuzzy inference system (FIS) model on household incomes data are the main focus of this study.

As increasing of size and complexity of data sets, the importance of data mining also has been increased. Data consist of many forms that need to be analyzed, processed and converted into information in order to inform, instruct, answer, and aid understanding of decision making. In other words, data mining is a process of

gaining knowledge from a big data with application of data mining tools. The increasing volume of data has made the data mining task as a computer-based method, which is a method used to process and to analyze the data (Cebeci and Yildiz, 2015).

Clustering is a famous method that is still being used in research because of its simplicity to compute groups based on similarity shared among the data. Cluster analysis is a formal study of methods using an algorithm to group the objects with similar characteristics into the same cluster. This method is sometimes known as unsupervised classification, where there is no information about the class labels or the number of clusters. There are many applications of clustering in the real world problems such as data documentation, image processing, pattern recognition and spatial data analysis. The goal of clustering is to discover new groups of interest among themselves and it usually can be applied for better understanding and utility on the new groups of interest. The best clustering technique minimizes the distance within the same group and maximizes the distance from the other clusters (Sarafis, 2005).

Particularly, *K*-means clustering was introduced by MacQueen in 1967 (MacQueen, 1967) and then it is popular for cluster analysis in data mining. *K*-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Although *K*-means clustering algorithm is simple and popular, *K*-means solutions are not guaranteed to be globally optimal due on the sum of square error (SSE) is the smallest compared to any other possible solution. It has a fundamental weakness of falling into local optima that depend on the randomly generated initial centroid values. An inappropriate choice of *k* clusters may yield poor results. Besides, the results might change when the analysis is conducted again.

Basically, *K*-means is applicable only when the mean is defined and it is very sensitive to outliers. Thus, it is not suitable for non-convex shapes. That is, the reason when performing *K*-means clustering on the data, it is crucial to run diagnostic checks for determining the number of clusters in the data set (Singh *et al,* 2013). Moreover, variants exist in the *K*-means clustering output due to the selection of the initial centre of *k* clusters, dissimilarity calculation and method to the calculated cluster means.

In this thesis, to make the clustering outputs more significant, clustering is used as a step to collect information before FIS is built. As many things in this world involve uncertainty or unclear boundaries, fuzzy logic is a mathematical approach that can be used to represent the vague or the uncertainty. The theory of fuzzy logic is based on the related graded membership and it has the ability to model uncertain or ambiguous data. This fuzzy method was introduced in 1965 by Lotfi Zadeh (Zadeh, 1965). In order to make suitable decisions for uncertainty, he proposed the set of membership which the membership value is "1" if it belongs to the set and "0" if it does not belong to the set. This membership idea was then extended to possess various "degree of membership" on the real continuous interval [0,1]. The fuzzy concept has contrast with the crisp set which requires the boundary to be precise (Sivanandam *et al*., 2006).

FIS is the process of formulating the mapping from a given input to the desired output by using fuzzy logic concept. The mapping provides a basis from which decisions can be made or patterns are discerned. Rules form is the basis of the fuzzy logic which is comprising the rule-based system originates from sources other than that of human experts and hence it is different from expert systems. It involves linguistic variables as antecedents and consequents. The antecedents express an inference or the inequality while the consequents are those that can be inferred as an output if the antecedent inequality is satisfied (Sivanandam *et al*., 2006). The number of rules increases exponentially with the dimension of the input space. The system uses IF-THEN rule-based system, which is

IF antecedent, THEN consequent.

FIS consists of fuzzification interface, a rule base, a database, a decision-making unit and defuzzification interface. As shown in Figure 1.1. Mamdani and Takagi-Sugeno systems are two examples of FIS that are widely used in various fields such as automatic control, data classification, decision analysis, expert systems, and computer vision (Sivanandam *et al*., 2006). Mamdani's method is among the first control systems built by using fuzzy set theory. It was proposed in 1975 by Ebrahim Mamdani (Mamdani, 1975) as an effort to control a steam engine and boiler combination by synthesizing a set of linguistic control rules that was obtained from experienced human operators. Mamdani's effort was based on the paper of Lotfi Zadeh's (1973) on fuzzy algorithms for complex systems and decision processes. Mamdani FIS expects the output of the membership functions to be fuzzy sets.

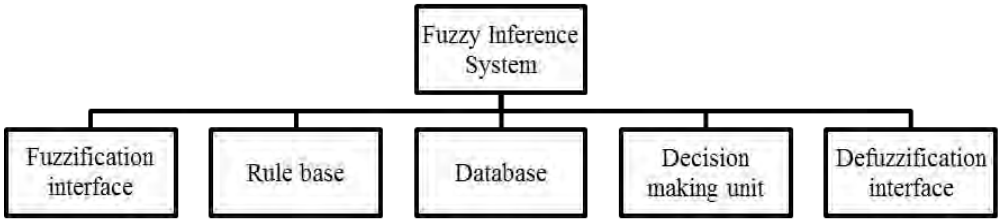

Figure 1.1: Fuzzy inference system structure

Meanwhile, the 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, 2012). The Takagi-Sugeno model was proposed by Takagi, Sugeno, and Kang which is known as an effort to formalize a system approach to generate fuzzy rules from an input-output data set. The format of Takagi-Sugeno fuzzy rule is given by

$$\text{IF } x \text{ is } A \text{ and } y \text{ is } B, \text{ THEN } z = f(x, y)$$

where A and B are fuzzy sets in the antecedents; $z = f(x, y)$ is a crisp function in consequent.

Both systems have advantages in modeling. Mamdani FIS is known as intuitive, widespread acceptance and suited with human inputs, whereas Takagi-Sugeno FIS is computationally efficient, works well with linear techniques, optimization and adaptive techniques, has guaranteed continuity of the output surface and well suited to mathematical analysis (Sivanandam *et al*., 2006). Thus, both of these methods will be used to test the efficiency of the best fuzzy system produced. Fuzzy models are evidently proven that they provide better solutions for complex problems. In this study, three shapes of membership function which are triangular, trapezoidal and Gaussian membership function used in the models to compare the efficiency of the models.

Furthermore, in this study, the data obtained are used to identify the group of household based on their incomes data provided. The groups of household income are important to identify the sources of incomes are varies. Thus, based on similarities exist in the data collected, the pattern of the groups formed can be identified. However, the range of each income to be considered as group members are unknown. By using the clustering result as preliminary knowledge, the prediction model of FIS will be built. The aim is to use the *K*-means clustering on the household

income data to produce a FIS model. The membership value for the FIS model is built using cluster central value. Based on the final results of the FIS model, a conclusion can be made on FIS that can be built based on non-fuzzy cluster outputs.

The models built are compared with the existing classification technique known as discriminant analysis. The percentage of the data correctly classified are recorded to compare the performance of the models built with discriminant analysis. By comparing the models with existing model, this research becomes more significant.

Figure 1.2 shows the procedure of achieving the aim of this study. The process of building household incomes prediction model involved data collection, *K*-means clustering process, the building of FIS, and computational of root mean square error (RMSE) value for each model. The best model selection is based on the RMSE value.

Figure 1.2: The process of building model for household income

1.2 Problem statement

The economic status of citizens in a country, which can be measured by carrying out the household incomes survey, is really important in making the income and consumption policies. Due on the sources of incomes of each household are varied, the clusters of the incomes are not easy to be determined. Since the similarities and differences on the household incomes reveal the behavior of the incomes data, the clustering of the income data gives a challenging task for the economists and the policy makers. The income data can be categorized into the same group once the similarities among the household incomes are measured. In such a way, the groups of the household could be found and the information gained can be used to build a prediction model for identifying the class of household incomes. However, the number of classes based on the household incomes are still unclear.

Moreover, Malaysia is a developing country, thus, finding the clusters of household incomes becomes an interesting issue because the groups of the cluster will show the economy of the population, which is divided based on their similarities. The range of every income data in each group could be very useful in building household incomes class prediction model. For this prediction model purpose, a preliminary knowledge is needed. However, the process of building models is quite complicated because outputs from clustering process are varies. The software used is usually required to rerun for the same value of the number of clusters or for different value of the number of clusters in which the results are compared for selecting the best result.

Based on clustering output, there is no strong information that could be obtained from the clustering process as the clustering is based on a comparison of objects and its variate but not on the estimation of the variate itself. The range for every data in the same group also does not clear. Thus, this unclear range would lead to difficulties for class decision. In addition, in economy sector, taxation system, rebate, financial incentive and bonus that are given by the government are made based on the citizens' income. For example, taxation should be applied to citizens by identifying which class can contribute better to Malaysia's economy. With this unclear range of incomes, setting up the policy to improve economy would not be success. Hence, a standardized range of household income class are very crucial and building a suitable prediction model will help the policy maker to discover the income class in advanced.

1.3 Objective of study

The following are the objectives of this study:

(a) To apply *K*-means clustering on household incomes data.

(b) To propose the prediction of household incomes classification using FIS.

(c) To analyze and validate the classification solution for household incomes with discriminant analysis.

1.4 Scope of study

This study focuses on building the household incomes model with *K*-means clustering FIS approach. The data used is from the Department of Statistics Malaysia (DOSM) provided by UTHM Micro Data which is handled by the Department of Mathematics and Statistics, Faculty of Applied Science and Technology. As this study intends to build fuzzy systems of incomes, only necessary variables, which are related to incomes of the citizens, are selected. Four incomes used are employment incomes, self-employment incomes, property incomes and current transfer received. However, the total number of data used is just a part of all data of Malaysians. The groups of citizens based on data provided will be obtained by using *K*-means clustering. Basically, the application of *K*-means clustering is as preliminary steps to gather information of the selected data before prediction model of household incomes is built. The value of centers of each cluster gained is used to build rules of FIS. Two types of fuzzy methods which are Mamdani and Takagi-Sugeno systems are used with different types of membership functions which are triangular, trapezoidal and Gaussian to find the best method to achieve the research objectives. Using these approaches, the clustering outputs are more meaningful which lead to building the prediction model of household incomes class. The percentage of data correctly classified of the models is compared with discriminant analysis output. The value of RMSE of the models is also identified.

1.5 Significance of study

Recent development in the field of data mining has led to a renewed interest in exploring the data. This study used household incomes data as the main focus to develop a class prediction model. The household incomes is an important component of citizens' economy status indicator. Thus, by building the prediction model, class of household incomes can be identified, so that it will be easier to identify which class of certain incomes belong to. Another significance of this study is by applying *K*-means clustering, the number of groups of household incomes that share similarities is managed to find out. Despite this, the model that is built using FIS approach helps to identify the prediction of classification of household incomes. The

understanding of household incomes classification is important in order to identify how each data is put in the same class based on the estimation of each variate that has been made using fuzzy rules. In order to identify the most accurate prediction model, the RMSE value is computed; the model with the smallest value of RMSE is concluded as the best model. By comparing the models with existing discriminant analysis, the models' performance in identifying data class would be proven. As discussed above, it shows that the efficiency of FIS to work on non-fuzzy clustering result to build class prediction model is successfully identified.

1.6 Thesis outlines

This thesis consists of five chapters including the introduction of the study. Chapter 1 focuses on an overall overview of research topics. The problem statement, objective, scope and significance of the study are discussed in this chapter.

In Chapter 2, the past research on clustering, fuzzy logic, and household incomes are reviewed. The past researches are very useful as guidelines to carry out this research. The methods that have been used also are identified to make a comparison for this research. Besides, the reviews are aimed to identify how wide the method and the same type of data have been used in the research field.

As moving to Chapter 3, it explains how each method is carried out to achieve research objectives starting from data collection data. Each method is explained using figures, flow charts and formulae provided to increase understanding of readers about the research. For *K*-means clustering method, the steps of how the clustering is carried out are shown and each distance measure used is listed out. Meanwhile, the process of building model using FIS is explained starting with building fuzzy rules, shape of membership function, type of FIS and computation of RMSE for model performance comparison. The reference of each method used is also stated in this section.

In Chapter 4, the results of the research are demonstrated. All the outputs from the research carried out are interpreted in this chapter. The *K*-means clustering outputs are given by using cluster figures and Silhouette index. The clusters centers then are used for building fuzzy rules. The rules view of each FIS model is shown in this chapter too. The RMSE value for each model is also computed and the

percentage of data correctly classified is compared with discriminant analysis. Finally, the best model can be identified.

For Chapter 5, the contributions of the research, limitations, and recommendations of the study are discussed. The contribution of the research state how the research is useful to the readers and future research direction. Meanwhile, the limitations highlight circumstances that need to face when the research is carried out; thus, the recommendations are the idea to improve the research to obtain better outcomes in future research. Finally, a summary of the study and a conclusion to the readers are made.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter is a review of previous studies regarding clustering and fuzzy inference system (FIS). There are many clustering techniques which will lead to different clustering results. A well-known *K*-means clustering is chosen as a technique to extract information from household incomes data before FIS is built. On this basis, a review of relevant studies on both clustering and fuzzy theory is further discussed.

2.2 Clustering as a data mining technique

Data mining is the process of gaining useful information from large data repositories to find unexpected relationships and to summarize the data in novel ways to make the data become understandable and useful to the data owner (David *et al.,* 2001). In data mining, the databases that always been used have millions of records and thousands of variables. The existing of too many predictor variables usually complicate the process to model a relationship with a response variable. Thus, data mining techniques are an approach to find understandable solutions (Larose, 2006).

The well-known data mining techniques are anomaly detection, association rule learning, classification, clustering and regression (Karaoussi and Bouhmala, 2012). The common data mining tasks are description and summarization, concept descriptions, segmentation, classification and case-based reasoning, prediction and dependency analysis (Han *et al*., 2001). The purpose of deploying data mining techniques is to discover crucial patterns from datasets and also provide capabilities to predict the outcome of a future observation. However, the relationships that exist in the data have the tendency to be unclear regarding the amounts of information are too big or the types of relationships are very difficult to imagine (Solarte, 2002). Data mining is an extension of statistical methods which are known to be a technique of increasing the productivity of people trying to build a predictive model (Two Crows, 1999).

Clustering analysis is one of the most frequently used data mining techniques (Solarte, 2002). Clustering is an unsupervised learning task with a purpose to make natural groupings based on similarity (Kruse *et al*., 2007). Clustering process will divide the data sets that have similarities into same cluster group whereas objects that belong to the other clusters are as dissimilar as possible. Cluster analysis is used to separate data elements into groups by maximizing the homogeneity within elements of clusters and heterogeneity between clusters (Hair *et al*., 1998). Clustering is also known as unsupervised learning algorithm because the actual number of clusters is unknown (Vermunt and Magidson, 2002). The way on how a cluster could be recognized is not entirely clear but one feature of the recognition process would appear to involve assessment of the relative distances between points (Everitt *et al*., 2011). By using clustering technique, the data reveals a consistent pattern which then is sorted into subsets that are easier to analyze (Two Crows, 1999).

Vijay, Mahajan and Kandwal (2012) stated that clustering technique has advantages in reducing the cost of collecting and labelling large set of sample patterns and to find the natural groups within data by the preprocessing technique. Andritsos (2002) in his research presented the state of the art in clustering technique from data mining point of view. The two main categories of clustering are partitional and hierarchical clustering algorithms. However, there are many methods that had emerged in cluster analysis. Its performance always depends on the size and dimensionality of data. Good clustering technique is said to have:

a) Ability to perform well with massive data
b) Ability to investigate single and mixture attributes
c) Could find arbitrary-shaped clusters
d) Minimum parameters needed
e) Ability to handle noise
f) Sensitivity to the order of input records
g) Ability to handle high dimension data
h) Interpretability and usability

Many of clustering techniques are proposed but the stability of these techniques is still not very clear even unknown. The effect to the output of these clustering algorithms is still not discovering if the input changes slightly. Besides, these techniques are based on co-occurrences of the data objects and do not deal with mixed attributes. Clustering is widely used for a variety of fields including data mining (Fayyad *et al*., 1996), statistical data analysis (Kaufman and Rousseeuw, 1989), and compression (Zhang *et al*., 1997).

Clustering can be divided into two main categories, which are hard clustering and soft clustering. In hard clustering, the object belongs to only one cluster while in soft clustering, the object belongs to two or more clusters depending on its membership value. Basically, there are many techniques of hard clustering analysis such as hierarchical and partitioning as suggested by Fraley and Raftery (1998), and density-based, model-based and grid-based suggested by Han and Kamber (2001).

K-means clustering was proposed by MacQueen (1967) and it is still been used by many researchers for many years due to its simplicity computations. *K*-means clustering is known as hard clustering since each data point belongs to one cluster only (Bora and Gupta, 2014). *K*-means algorithm is based on the centre of the associated cluster (Faber, 1994). By using *K*-means clustering, the data is put into a homogenous cluster which means that the groups have identical characteristics; in terms of intra-cluster similarity (Anderberg, 1973; Bora and Gupta, 2014).

K-modes clustering uses similar concept of *K*-means but removes the limitation of numeric data (Khan and Kant, 2007). *K*-modes clustering is used for categorical data which adjusting *K*-means method by substitute Euclidean distance metric with simple matching dissimilarity measure. This technique uses modes to represent cluster centres and updating modes with the most frequent categorical

REFERENCES

Abbod, M.F. & Al-Shammaa, M. (2012). Automatic Generation of Fuzzy Classification Rules from Data. *Proceedings of the International Conference on Neural Networks- Fuzzy Systems,* 74-80

Abonyi, J., & Szeifert, F. (2007). Fuzzy Clustering for the Identification of Takagi-Sugeno Fuzzy Models of MIMO Dynamical Systems. 1-17. Retrieved from https://www.researchgate.net/publication/2869488

Abraham, A. (2004). Business Intelligence From Web Usage Mining. *Journal of Information & Knowledge Management,* 2(4),375-390. Retrieved from https://doi.og/10.1142/S0219649203000565

Acqua, G.D. & Abbondanti, R.L.F.(2003). Adaptive Neuro Fuzzy Inference System For Highway Accidents Analysis. University of Naples "Federico II".

Ahmad, Z. & Ejaz, Z. (2011). Classification of Households With Respect to Poverty by Using Cluster Analysis. *Proceeding of International Conference on Complex System,* 21, 369-381.

Alayande, S.A, & Adekunle, B.K. (2015). An Overview and Application of Discriminant Analysis in Data Analysis. *Journal of Mathematics,* 11(1), 12-15.

Ali, O.A.M., Ali, A.Y. & Sumait, B.S. (2015). Comparison Between The Effects of Different Types of Membership Functions On Fuzzy Logic Controller Performance. *International Journal of Emerging Engineering Research and Technology, 3(3),76-83.*

Anderberg & Michael, R (1973). Cluster Analysis for Applications. New York: Academic Press.

Alsabti, K., Ranka, S., & Singh, V. (1998). An Efficient K-means Clustering Algorithm. *IPPS/SPDP Workshop on High Performance Data Mining. IEEE Computer Society Press,* 881-892.

Andritsos, P. (2002). Data Clustering Techniques. University of Toronto: Rapport Technique.

Bandyopadhyay, S., & Saha, S. (2013). Unsupervised Classification. Springer Berlin Heidelberg.

Bhatia, M.P.S & Khurana, D. (2013). Analysis of Initial Centers for *K*-means Clustering Algorithm. *International Journal of Computer Applications,* 71, 9-12.

Bezdek, J.C. (1981). *Pattern Recognition with Fuzzy Objective Function Algorithms.* New York, Plenum Press.

Bogale, A., Hagedorn, K., & Korf, B. (2005). Determinants of Poverty in Rural *Ethiophia. Journal of Agriculture*, 44 (2), 101-120.

Bora,D.J., & Gupta, A.K. (2014). A Comparative Study Between Fuzzy Clustering Algorithm and Hard Clustering Algorithm. *International Journal of Computer Trends and Technology,* 10, 108-113.

Bora, D.J. & Gupta, A.K. (2014). Effect of Different Distance Measure on the Performance of *K*-means Algorithm : An Experimental Study in Matlab. *International Journal of Computer Science and Information Technologies,* 5, 2501-2506.

Bradley, P.S., Mangasarian, O.L., & Street, W.N. (1998) Clustering via Concave Minimization. *Advances in Neural Information Processing Systems,* 9, 368-374.

Buckley, J.J. & Hayashi, Y. (1994). Fuzzy Neural Networks: A Survey. *Fuzzy Sets and Systems,* 66(1), 1-13.

Calinski, R.B., & Harabasz, J. (1974). A Dendrite Method for Cluster Analysis. *Communications in Statistics- Simulation and Computation,* 3(1), 1-27.

Cebeci, Z., & Yildiz, F. (2015). Comparison of K-Means and Fuzzy C-Means Algorithms on Different Cluster Structures. *Journal of Agricultural Informatics,* 3,13-23.

Chang, C.T, Lai, J.Z.C., & Jeng, M.D. (2011). A Fuzzy K-means Clustering Algorithm Using Cluster Center Displacement. *Journal Of Information Science And Engineering ,* 27,995-1009.

Davies, D., & Bouldin, D. (1979). A Cluster Separation Measure. *IEEE Pattern Analysis and Machine Intelligence,* 1(2), 224-227.

Deborah, L.J., Baskaran, R., & Kannan, A. (2010). A Survey on Internal Validity Measure for Cluster Validation. *International Journal of Computer Science & Engineering Survey,* 1(2),85-102.

Devi, M.K. & Rani, M.U.(2013). Fuzzy Inference System and Its Application. *International Journal of Engineering Sciences Research,* 4, 1248-1250.

Diaz, B. & Morillas, A. (2008). Robust Statistics and Fuzzy Industrial Clustering, Forging the New Frontiers : Fuzzy Pioneers II. 219 Springer-Verlag Berlin Heidelberg.

Dunn, J.C. (1974). Well Separated Clusters and Optimal Fuzzy Partitions. *J.Cybern.,* 4(3), 95-104.

Dziopa, T. (2016). Clustering Validity Indices Evaluation with Regards to Semantic Homogeneity. *Position Papers of the Federated Conference on Computer Science and Information Systems,*9, 3-9.

Er, M.J. & Zhou, Y. (2008). Automatic Generation of Fuzzy Inference Systems via Unsupervised Learning. *Neural Network,* 21(10), 1556-1566.

Everitt,B.S., Landau, S., & Leese, M. (2001). Cluster Analysis. Arnold Publishers, London, fourth edition.

Fan, J.L., Zhen, W.Z., & Xi, W.X. (2003). Suppressed Fuzzy C-Means Clustering Algorithm. *Pattern Recognition Letters,* 24(9-10): 1607-1612.

Faber,V. (1994). Clustering And The Continuous K-means Algorithm. *Los Alamos Science*, 22, 138-144.

Farahbod, F. & Eftekhari,M. (2013). A New Clustering-Based Approach For Modelling Fuzzy Rule-Based Classification Systems. *International of Science and Technology, Transactions of Electrical Engineering,* 37, 67-77.

Fayyad, U., Shapiro, G.P., & Smyth, P. (1996). Data Mining to Knowledge Discovery in Database. *IEEE Expert: Intelligent Systems and Their Applications,* 11(5), 20-25.

Finley,T., & Joachims,T. (2008). Supervised K-means Clustering. *Computing and Information Science Technical Reports.*

Forgey, E.W. (1965). Cluster Analysis of Multivariate Data : Efficiency Versus Inter-Pretability of Classifications, Biometric Society Meeting, California. *Abstract in Biometrics 21 (1965) 768.*

Garey, M., & Johnson, D. (1979). Computer and Intractability- A Guide to the Theory of NP-Completeness, Freeman, New York.

Gore Jr., P.A., 2000. Cluster Analysis. In Handbook of Applied Multivariate Statistics and Mathematical Modelling. Academic press.

Gowda,K.C & Diday,E. (1992). Symbolic Clustering Using A New Dissimilarity Measure. *Systems, Man and Cybernetics, IEEE Transactions,22, 368-378.*

Guillaume, S (2001). Designing Fuzzy Inference Systems from Data : An Interpretability-Oriented Review. *IEEE Transactions on Fuzzy Systems,* 9 (3), 426- 443.

Hair,J.F.jr., Anderson,R.E., Tantham,R.L., & Black,W.C. (1998). Multivariate Data Analysis. Prentice Hall.

Han, J. & Kamber, M. (2001). *Data Mining: Concepts and Techniques. USA:* Morgan Kaufmann Publishers.

Household Income and Expenditure (1987). Retrieved from: https://www.dosm.gov.my/v1/index.php?r=column/cone&menu_id=cUp6NlNndGlaQkZhK0gwYUMyWFRxdz09

Huang,Z. (1998). Extensions To The K-Means Algorithm For Clustering Large Data Sets With Categorical Values. *Data Mining and Knowledge Discovery,* 2, 283–304.

Huberty, C.J. (1994). *Applied Discriminant Analysis.* New York: Wiley.

Huberty, C.J. & Hussein, M.K. (2002). Some Problems in Reporting Use of Discriminant Analysis. *Journal of Experimental Education*, 71, 177-191.

Jamsandekar,S.S. & Mudholkar, R.R. (2014). Fuzzy Classification System by Self Generated Membership Function Using Clustering Technique. *BIJIT-BVICAM's International Journal of Information Technology,* 6(1), 142-152.

Kalpana, M. & Kumar, A.V. (2011). Fuzzy Expert System for Diabetes using Fuzzy Verdict Mechanism. *International Journal Advanced Networking and Applications,* 3(2), 1128-1134.

Kanungo,T., Mount, D.M., Netanyahu, N.S., Piatko,C., Silverman,R., & Wu, A.Y.(2000). An Efficient k-Means Clustering Algorithm: Analysis and Implementation. *Proceedings of the Sixteenth ACM Symposium on Computational Geometry,* 100-109

Karoussi, E, & Bouhmala, A.(2012). Data Mining K-Clustering Problem. University of Agder: Master's Thesis.

Kaufman, L., & Rousseeuw, P. (1990). Finding Groups in Data: An Introduction to Cluster Analysis. New York, John Wiley & Sons.

Kaur,A. & Kaur, A. (2012). Comparison of Mamdani-Type and Sugeno-Type Fuzzy Inference System for Air Conditioning System. *International Journal of Soft Computing and Engineering (IJSCE),* 2(2), 2231-2307.

Khan, S., & Kant, S. (2007). Computation of Initial Modes for K-modes Clustering Algorithm Using Evidence Accumulation. *IJCAI International Joint Conference on Artificial Intelligence,* 2784-2789.

Khare, Y.B. (2011). Development of Medical Inferencing System Using Data Clustering. Thapar University: Master's Thesis.

King, R.S. (2015). *Cluster Analysis and Data Mining. An Introduction.* New Delhi : Mercury Learning and Information.

Kruse, R., Doring, C., & Lesot, M.J. (2007). A*dvances in Fuzzy Clustering and Its Application.* USA: John Wiley & Sons, Ltd.

Larose, D.T. (2006). *Data Mining Methods And Models.* United States of America, John Wiley& Sons.

Lee, C.S. (2009). *A Framework of Adaptive T-S Type Rough-Fuzzy Inference Systems.* University of Western Australia: Ph.D. Thesis

Li, H., & Chen, C.L.P. (2000). *Fuzzy Neural Intelligent Systems: Mathematical Foundation and Applications in Engineering.* USA: CRC Press.

Liu, Y., Li, Z., Xiong, H., Gao, X., & Wu, J. (2010). Understanding of Internal Clustering Validation Measures. *IEEE International Conference on Data Mining,* 911-916

MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. *Proceeding Symposium on Mathematical Statistics and Probability,* 233, 281–97. Retrieve from http://dx.doi.org/citeulike-article-id:6083430

Mamdani, E.H. & Assilian, S. (1975). An Experiment in linguistics synthesis with a fuzzy logic controller. *International Journal Man-Machine Studies*, 7, 1-13. DOI: 10.1016/S0020-7373(75)80002-2.

Mamdani, E.H. (1977). Applications of Fuzzy Logic to Approximate Reasoning Using Linguistic Synthesis. *IEEE Trans.Comput.,*26(12), 1182-1191.

Mandal, S.N., Choudhury, J.P., & Chaudhuri, S.R.B. (2012). In Search of Suitable Fuzzy Membership Function of Time Series Data. *International Journal of Computer Science Issues.* 9(3), 293-302.

Mansoori, E.G. FRBC: A Fuzzy Rule-Based Clustering Algorithm. *IEEE Transactions on Fuzzy Systems.* 2011. 19 (5), 960-971.

Mooi, E., & Sarstedt, M. (2001). *A Concise Guide to Market Research.* Verlag Berlin Heidelberg: Springer.

Moreno, J.E., Castillo, O., Castro, J.R., Martinez, L.G., & Melin, P. (2007). Data Mining for Extraction of Fuzzy IF-THEN Rules using Mamdani and Takagi-Sugeno-Kang FIS. *Engineering Letters.* 15(1), 82-88.

Narayan, D., & Prichett, L. (1997). Cents and Socialibility: Household Income and Social Capital in Rural Tanzania. *World Bank*, 47(4), 871-897.

Nayak, G.K., Narayanan, S.J., & Paramasivan, I. (2013). Development and Comparative Analysis Of Fuzzy Inference System for Predicting Customer Buying Behavior. *International Journal of Engineering and Technology (IJET),* 5,4093-4108.

Nicklaus, C. (2015). *The Effect of Household Income on Household Consumption in China.* Lund University; Master's Thesis.

Pandit, S., & Gupta, S. (2011). A Comparative Study on Distance Measuring Approaches for Clustering. *International Journal of Research in Computer Science,* 2(1), 29-31. *Doi: 10.7815/ijorcs.21.2011.011.*

Parmar, K.J. (2016). *Cluster Validation of Whan Galaxy Classification Using A Novel Approach to External Cluster Validation*. University of Houston ; Master's Thesis.

Patidar, A.K., Agrawal, J., & Mishra, N. Analysis of Different Similarity Measure Functions and their Impacts on Shared Nearest Neighbor Clustering Approach. *Journal of Computer Applications,* 40(16): 1-5.

Pham, D.T., Dimov, S.S., & Nguyen, C.D. (2004). An Incremental *K*-Means Algorithm. *Proceedings of the Institution of Mechanical Engineers,* 218 (7), 783-795.

Pyryt, M.C. (2004). Pegnato Revisited: Using Discriminant Analysis to Identify Gifted Children. *Psychology Science,* 46 (3), 342-347.

Rahman, R. (2006). *Access to Education and Employment: Implications for Poverty in Bangladesh.* Programme for Research on Chronic Poverty in Bangladesh (PRCPB), Working Paper No. 14, World Bank Report, 2009.

Reardon, T., Fall, A.A., Kelly, V., Matlon, C.D., Hopkins., J., & Badiane, O. (1993). Agriculture-Led Income Diversification In The West African Semi-Arid Tropics: Nature, Distribution, and Importance of Production Linkage Activities. *International Conference of African Economic Issues.*

Ross, T.J. (2010). *Fuzzy Logic with Engineering Applications.* USA: John Wiley and Sons.

Rousseeuw, P.J. (1986). Silhouettes: A Graphical Aid To The Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

Sarafis, I. (2005). Data Mining Clustering of High Dimensional Databases With Evolutionary Algorithms. Heriot-Watt University : Doctor of Philosophy .

Sharma,S.C (1996). *Applied Multivariate Techniques.* USA: John Wiley and Sons.

Shinde, S.A., & Solanki, S.S. (2012). A Fuzzy Rule Based Clustering Development Novel. *International Journal of Science and Research,* 3(12), 2012-2015.

Shinde, S.V., & Kulkarni, U.V. (2014). Modified Fuzzy Hyperline-Segment Neural Network for Classification with Mixed Attributes. *International Conference on Computing, Communication and Networking Technologies,* 5, 1-7.

Silviu, B. (2001). Fuzzy Clustering. *Semantic Scholar.* Retrieved from https://www.semanticscholar.org/paper/Fuzzy-Clustering Silviu/13e43b4710620fb3e6e303d612f541e8e74777d6?tab=relatedPapers

Sivarathri, S. & Govardhan, A. (2014). Experiment On Hypothesis "Fuzzy K-Means is Better Than K-Means For Clustering. *International Journal of Data Mining & Knowledge Management Process,* 4(5): 21-34.

Solarte, J. (2002). A Proposed Data Mining Methodology and Its Application To Its Industrial Engineering. University of Tennessee : Master's Thesis.

Sutherland, H., Taylor, R., & Gomulka, J. (2001). Combining Houesehold Income and Expenditure Data in Policy Simulations. University of Cambridge. *Review of Income and Wealth,* 4, 517-536.

Takagi, T. & Sugeno, M. (1985). Fuzzy Identification of Systems and Its Applications to Modelling and Control. *IEEE Transactions on Systems, Man and Cybernatics,* 15(1),116-132.

Tosi, S., Casolari, S., & Colajanni, M. (2013). Data Clustering Based on Correlation Analysis Applied to Highly Variable Domains. *Computer Networks,* 57(15): 3025-3038.

Tsekouras, G.E., Pavlogeorgatos, G., & Kalloniatis, C. A Fuzzy Clustering Algorithm for Generating Fuzzy Rules from Numerical Data. Retrieved from http://www.academia.edu/2671888/

Two-Crows. (1999). *Introduction to Data Mining and Knowledge Discovery.* USA: Two Crows Corporation.

Varjonen J, Aalto K (2005) *Household production and consumption in Finland 2001.* Household Satellite Account. Statistics Finland and National Consumer Research Centre.

Vermunt, J.K., & Magidson, J. (2002). *Latent Class Cluster Analysis. In Applied Latent Class Analysis.* Cambridge: Cambridge University Press.

Vijay, R., Mahajan, P., & Kandwal, R. (2012). Hamming Distance based Clustering Algorithm. *International Journal of Information Retrieval Research,* 2, 11-20.

Wagstaff,K., Cardie,C., Rogers,s., & Schroedl,S. (2001). Constrained K-Means Clustering with Background Knowledge. *Proceedings of the Eighteenth International Conference on Machine Learning,* 577-584.

Xie, X.L., & Beni, G.. A Validity Measure for Fuzzy Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 13(8), 841-847.

Zafari, A. (2014). Developing A Fuzzy Inference System by Using Genetic Algorithms. University of Twente: Master's Thesis.

Zadeh, L.A. (1965). Fuzzy Sets. *Information Control,* 8, 338-353.