TELKOMNIKA, Vol.16, No.6, December 2018, pp.2725~2732
ISSN: 1693-6930, accredited First Grade by Kemenristekdikti, Decree No: 21/E/KPT/2018
DOI: 10.12928/TELKOMNIKA.v16i6.9478

Fuzzy Rule-based Classification Systems for the Gender Prediction from Handwriting

Lala Septem Riza*[1], Aldi Zainafif[2], Rasim[3], Shah Nazir[4]
[1,2,3]Department of Computer Science Education, Universitas Pendidikan Indonesia
Setibudhi St. 229, Bandung, Indonesia
[4]Department of Computer Science, University of Swabi, Swabi, Pakistan
*Corresponding author, e-mail: lala.s.riza@upi.edu

Abstract

The handwriting is an object that can describe information about the author implicitly. For example, it is able to predict the gender. Recently, the gender prediction based on handwriting becomes an interesting research. Even in 2013, an competition for prediction gender from handwriting has been held by Kaggle. However, the accuracies of current approaches are relatively low. So, in this study, we attempt to implement Fuzzy Rule-Based Classification Systems (FRBCSs) for gender predictions from handwriting. Three stages are conducted to achieve the objective, as follows: defining some features based on Graphology Techniques (e.g., pressure, height, and margin on writing), collecting real datasets, processing on digital images (i.e., image segmentation, projection profiles, and margin calculation, etc.), and implementing FRBCSs. The implemented algorithm based on FRBCSs in this research is Chi's Algorithm, which is a method based on Fuzzy Logic for classification tasks. Moreover, some experiments and analysis, involving 75 respondents consisting of 36 males and 39 females, have been done to validate the proposed model. From the simulations, the classification rate obtained is 76%. Besides improving the accuracy rate, the proposed model can provide an understandable model by utilizing fuzzy rule-based systems.

Keywords*: fuzzy sets, fuzzy rule-based systems, gender prediction, graphology, image processing*

1. Introduction

The handwriting is one of the objects in research used for focusing on discovering private information about its authors. In handwriting, a scratch written by someone will be greatly influenced by the emotional, health, and past experience in itself [1]. It means that a person's handwriting can determine characteristics inside the writing, such as the level of emotion, confidence, and the gender [2]. Based on research conducted by Weisberg et al. [3], for example, a woman has the openness, curiosity, empathy, and emotion that is higher than men.

In 1897, the term "graphology" was introduced by Michon [1] in Paris by merging two Greek words: *graphein* meaning write and *logos* for science. At that time, he analyzed handwriting scientifically. Then, Jamin, who is one of Michon's students, continued his work by dividing the writing into seven fundamental aspects: speed, pressure, form, dimension, continuity, direction, and order [4]. By these techniques, this research also revealed health issues, morality and past experiences to hidden talents, and mental problems. Thus, it is possible that some aspects of graphology to be a parameter determining sex based on handwriting.

In 2013, research on gender prediction based on handwriting to be very interesting to discuss. This is evidenced by the competition, organized by Kaggle and attended by 194 teams from academia and industry, for gender prediction from images of handwriting [5]. Competitors are required to construct a model for the gender prediction by learning from datasets provided by the organizers, and then to make predictions on new data. This activity has been making a motivation in research for determining gender from handwriting. Currently, some researches on gender prediction based handwriting have been done. Some of them are done by Liwicki et al. [6], Maadeed and Hassaine [7], and Bradley [2]. In their studies, they mostly used some aspects in graphology, such as pressing a pen, writing curvature, etc., Furthermore, it can be seen from

Received April 3, 2018; Revised September 28, 2018; Accepted October 29, 2018

these studies that the classification rates are between 54% and 74%. So, the accuracy is still relatively low.

Therefore, this research is aimed to construct a model used for predicting a gender from handwriting by using Fuzzy Rule-Based Classification Systems (FRBCSs). The systems are proposed by utilizing fuzzy concepts to construct a set of fuzzy rules for dealing with classification tasks. In this case, we consider an algorithm included in FRBCSs, which is Chi's Algorithm. Basically, the proposed model contains three stages as follows: (i) we define some features based on the graphology techniques (e.g., pressure, height, slope, and margin on writing), (ii) some real datasets represented in images of handwriting are gathered, (iii) some processes on these images should be done, such as image segmentation, projection profile, and margin calculation, and (iv) finally we implement FRBCSs to construct a model represented by a set of fuzzy rules.

This paper is organized as follows section 2 introduces to fuzzy set theory and FRBCSs. Then, section 3 describes the proposed model for gender prediction from handwriting. We illustrate an experimental study in section 4 that is followed by its results and discussion in section 5. Finally, section 6 concludes the paper.

2. Fuzzy Rule-Based Classification Systems (FRBCSs)

Before explaining about FRBCSs, firstly this section will discuss fuzzy set theory. It was proposed as a generalization of classical set theory to express sets by introducing degrees of membership of elements in sets [8]. Therefore, the main difference between crisp set from fuzzy set is that while the first just has two values for membership: a member and non-member, the latter allows to the degree of membership between zero and one. Mostly, we express this membership by membership functions, such as trapezoid, triangular, Gaussian, etc. Moreover, the theory effectively expresses the model of a human expert knowledge because of the concept of linguistic variables represented by fuzzy sets. For example, we define that the imprecise value of the object a_1 is "cold", and its degree of the membership function μ is 0.7. This statement can be interpreted that a_1 can be included in a fuzzy value expressed by "cold" with the membership of the object in the fuzzy label is 0.7.

An implementation of fuzzy sets is fuzzy rule-based systems (FRBSs), which are a variant of rule-based systems that are usually written in the IF-THEN form. In a general form, we have "IF X_1 is A THEN Y is B", where X_1 and Y are input and output variables with A and B are fuzzy sets. It can be seen that instead of involving numerical values, we represent knowledge in a fuzzy rule involving fuzzy sets.

Furthermore, FRBCSs are a variant of FRBSs that are used for dealing with classification tasks. A main property of FRBCSs is that the consequent part is represented by a class C_j from the pre-specified class set $C = \{C_1, C_2, \ldots, C_M\}$ whereas the antecedent part is kept in linguistic variables [9]. For example, we are trying to select mangos according to their quality. So, we have following variables and their linguistic values to classify a mango fruit quality: dimension = {*small*, *medium*, *large*}; weight = {*light*, *medium*, *heavy*}; color intensity = {*lighter*, *neutral*, *darker*}. Moreover, we have two decision values on the output variable: *selected* and *rejected*. In a specific case, we can take a fuzzy rule, as follows:

> IF dimension is *large* and weight is *medium* and color intensity is *neutral* THEN decision is *selected*.

so, it can be seen that the fuzzy rules can be understood and interpreted by human.

The next question that should be addressed is how to construct fuzzy rules from data training. Basically, there are two ways to construct a model in FRBCSs: by human experts and learning from data by using machine-learning methods. This research focuses on the second one, where the architecture showing the processes is illustrated in Figure 1.

It can be seen that there are two stages, namely learning and prediction, in constructing an FRBCS model. On the learning step, we firstly need to supply data training. Then, we perform a learning method that mostly involves the following two aspects: structure identification and parameter estimation [10, 11]. In the structure identification, a rule base corresponding to pairs of input and output variables is determined. Then, in the parameter estimation, the optimal values of the membership function are obtained. It should be noted that these processes can be

done sequentially or simultaneously. In this research, we consider Chi's Algorithm as the learning method [9]. Basically, it is an extension of Wang Mendel's Algorithm [12], which is a learning method using the space partition approach to build FRBSs.

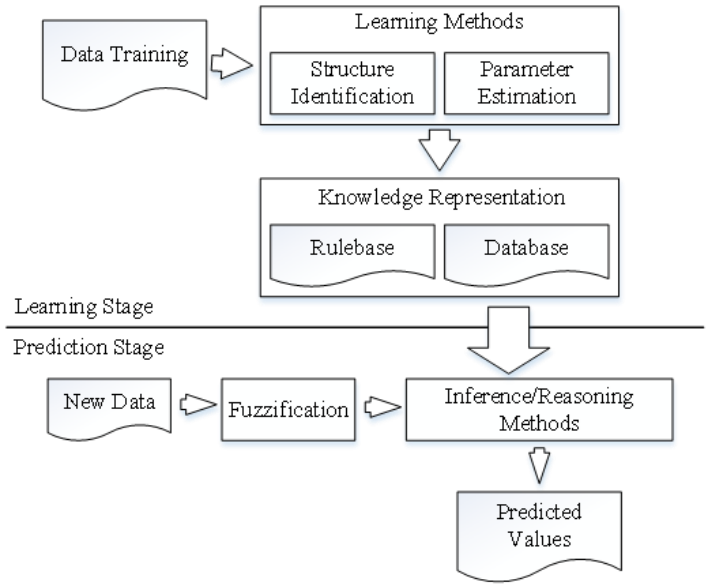

Figure 1. Learning and prediction stages on FRBCSs

For the learning step, the following steps are should be done:

1) Divide equally ranges of values of the input and output variables into fuzzy regions regarding defined linguistic values as the database. For example, we define the range of the mango weight between 0.0 kg and 1.6 kg. Then, the following linguistic terms are used: "*light"*, "*medium"*, and "*heavy"*. It means the number of linguistic terms is 3. So, we can calculate the fuzzy regions as the following intervals: [0, 0.8], [0.4, 1.2], and [0.8, 1.6] for "light", "medium", and "heavy", respectively. Now we can construct triangular membership functions that can be defined by the corner points. For example, in the "medium" value, we can construct a triangle with the corner points as a = 0.4, b = 0.8, and c = 1.2 where b is the middle point with the membership function degree of 1.
2) Built a set of fuzzy rules representing the training data by implementing the database from Step 1. So, in this step, degrees of membership function can be calculated for all values in the training data, and choose a linguistic value based on its maximal degree for the antecedent part. It should be noted that the consequent part is a categorical value/class.
3) Calculate a degree for each fuzzy rule. It is determined by fuzzy set operators (e.g., product) over degrees of membership functions in the antecedent part.
4) Obtain a final rule base after omitting redundant rules. Moreover, we can delete a redundant rule by considering its degree.

After accomplishing these processes on learning methods, an FRBCS model representing knowledge over the data have been constructed. Basically, the model contains two parts: rule base and database. A rule base is the same as a set of fuzzy rules whereas a list of fuzzy labels, membership functions, and their values are collected into database. Then, this model will be used on the next stage, which is prediction.

In the prediction part, after supplying a set of new data, we perform fuzzification, which is a process for transforming from crisp sets into fuzzy sets (i.e., represented by linguistic labels and their degrees of memberships) [10]. Then, by using the FRBCS model obtained from the learning step, inference should be done to predict values of new data. A complete explanation regarding these processes can be found in e.g., [13]. Moreover, implementations of fuzzy logic have been done by many researches. For example, the study of [14] utilized fuzzy logic for optimizing in a lighting control system; a fuzzy logic classification has been conducted to classify

the queuing incoming packet [15]; and the maximization the power output of solar panels can be also achieve by performing fuzzy logic controller [16].

3. The Model Construction for Gender Prediction

Figure 2 shows the model proposed to predict gender from handwriting in this research. It can be seen that basically we can divide it into 5 big processes as follows:

1) Designing features for generating data training. There are two following steps that should be done:
   a) Determining features used for input parameters on data training. In this case, we consider some features based on graphology, as follows: pressure on writing, height of letters and words, maximum height on sentences, center line on sentences, and margin.
   b) After designing the features, we need to choose letters, words, sentences, and paragraphs that are used as samples in this research. The following is a list of them:
      - Letters: {“a”, “d”, “e”, “g”, “k”, “m”, “t”}.
      - Words: {“Bilingual”, “Orange”, “Toblerone”}.
      - Sentences: {“The quick brown fox over the lazy dogs”, “A camel's tailbone is eaten by a lion”}.
      - Paragraph: {“A giraffe has steel pierced ears. Others chat with flying monkeys. Zebra peeks at x-rays with philosophy. Rhinos are finally angry at all.”}

      It should be noted that we choose these letters, words, sentences, and paragraphs since they are believed to represent any characteristics in graphology. So, by combining with the features and the chosen letters, words, and sentences, we obtain 49 parameters and 1 output variable having two values: male or female.
2) Data collection. In this step, two steps should be done as follows:
   a) Creating form for questioner. Basically, the respondents just need to re-write letters, words, and sentences indicated in the boxes with handwriting.
   b) Then, we delivered the questioners to some respondents. After a week, we obtained 75 data containing handwriting with their genders.
3) Image processing on the data. In order to obtain the features as determined previously, some standard techniques on image procession [17] should be done. For example, we need to perform image segmentation that is used for dividing an image into multiple segments (i.e., sets of pixels). In this case, we are using Otsu Threshold [18]. Then, we need to calculate height of letters and words, center line of words, pressure on letters and words, and margin of a paragraph. These features can be obtained by utilizing the projection profile technique [19]. Finally, the output of these processes is data that are used for training and testing on the experiments.
4) On the other hand, we also implement Chi’s Algorithm based on FRBCS, where it has been explained on the previous section.
5) To validate and test the proposed model, we perform some experiments. These are generated by conducting 5-fold cross validation. The complete explanation can be found on the next section.

4. Experimental Study

In this section we illustrate three important aspects that are used for experiments, as follows: collected datasets, cross validation, some parameters on learning with FRBCSs.

4.1. Datasets

As we mentioned previously, in this experiment we have collected image data from 75 respondents consisting of 36 males and 39 females. Then, before doing image processing as illustrated in Figure 2, we scan the data to get digital images. After calculating with image processing techniques, we obtain numerical data as shown in Table 1. These values are generated by techniques in image processing mentioned in the model.

So, it can be seen that the combinations between the defined features (height, centre line, pressure, and margin) and the chosen letters, words, and sentences generate 49 input parameters, such as X1 represents pressure on the letter “a”, X8 is pressure on the word “Bilingual”, X14 is height of the word “Bilingual”, and X49 is the left margin of the paragraph.

Actually, these attributes are generated by image processing on data training. The last column, which is Label, contains “Ma” and “Fe” representing the gender male and female.

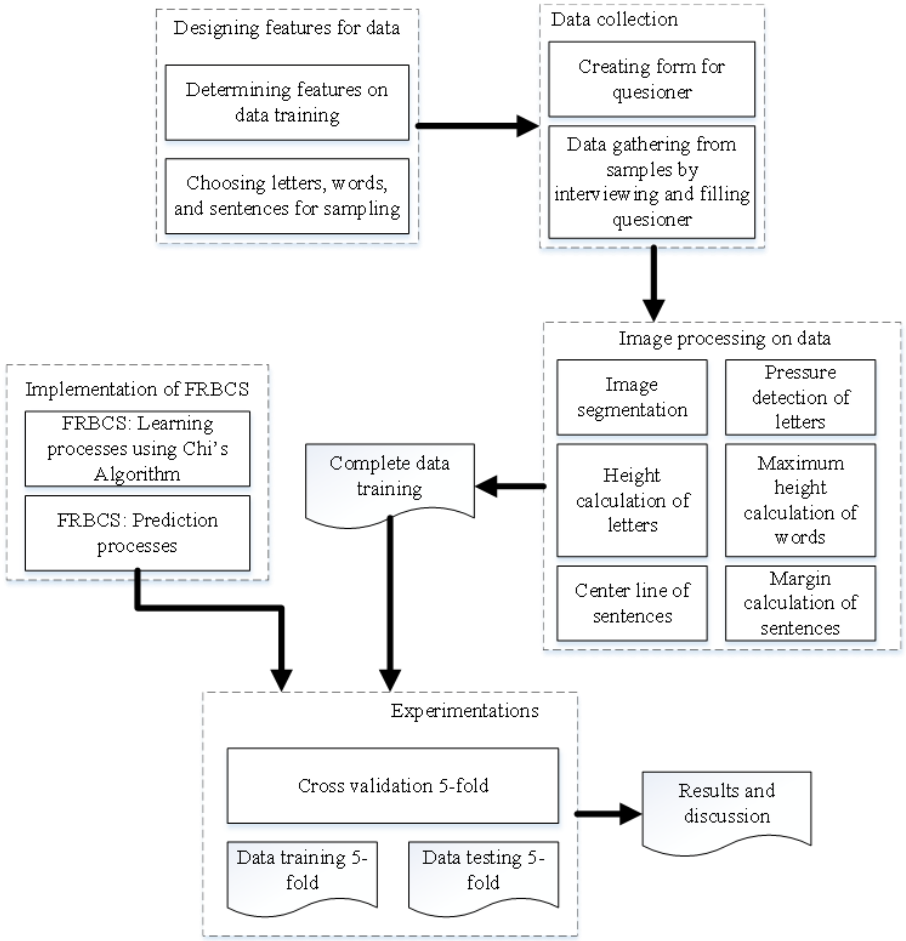

Figure 2. Research design on gender prediction using FRBCS and graphology techniques

Table 1. A Part of Data Training Containing 75 Samples with 49 Input Parameters

No	X_1	X_2	X_3	X_4	…	X_{46}	X_{47}	X_{48}	X_{49}	Label
1	0.635	0.754	0.586	0.648	...	7	0	46	22	Ma
2	0.486	0.661	0.797	0.958	...	0	283	116	33	Ma
3	1.130	1.566	1.283	1.776	...	12	62	183	18	Ma
4	0.479	0.599	0.938	0.670	...	16	85	281	22	Ma
…	...	...	...	...	...	...	...	...	...	...
74	0.975	0.800	0.790	1.096	...	3	14	162	59	Fe
75	0.692	0.653	0.887	0.876	...	40	269	71	53	Fe

4.2. Cross Validation

Before doing the experiment, the data are arranged into training and testing datasets by using the cross validation 5-fold [20]. So, we obtain 5 pairs of data training and testing, where 80% data is for training and the rest is for testing. In other words, we will simulate the experiments for 5 times. For each simulation, we then calculate the accuracy, and finally the average of them can be obtained.

4.3. The Learning Stage Using FRBCS

In order to execute Chi’s Algorithm, there are some parameters that should be defined, as follows: numbers of fuzzy labels/linguistics values and types of membership functions. In this

research, we set the numbers of fuzzy labels to be 3 and the membership function type to be triangular. So, if we assume minimum and maximum values of the parameter X_1 to be 0.29 and 5.8, we can visualize degrees of membership function as illustrated in Figure 3. It can be seen that the domain space of X_1 is divided by 3 fuzzy labels with the triangular membership function equally. In this case, the fuzzy labels are represented by the general codes "S1", "M", "B1". Of course, we can change them to be "Small", "Medium", and "Big". It should be noted that the $m(X_1)$ represents the degree of membership function that has a value between 0 and 1.

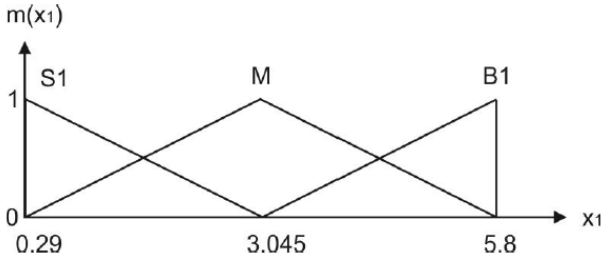

Figure 3. Membership function and their degree of the parameter X_1 with 3 fuzzy labels

5. Results and Discussion

As we mentioned in Figure 1, the model constructed by FRBCSs contains two parts: database and rule base. While database can be illustrated as in Figure 3 for all parameters, rule base is a set of fuzzy rules that generated by learning from data training. For example, in the first simulation of 5-fold cross validation, we obtained 60 fuzzy rules generated on the learning step, as shown in Table 2. As we mentioned previously, these rules are produced by Chi's algorithm including structure identification and parameter estimation.

Table 2. Some Examples of Fuzzy Rules Generated by FRBCS

Index Rule	X_1	X_2	X_3	X_4	…	X_{47}	X_{48}	X_{49}	Label
1st	S1	M	M	M	…	M	M	M	Ma
2nd	S1	S1	S1	S1	…	M	M	M	Fe
…	…	…	…	…	…	…	…	…	…
60th	S1	S1	M	S1	…	S1	M	M	Ma

For example, on the first rule we obtain the following fuzzy rule:

> IF X_1 is S1 and X_2 is M and X_3 is M and X_4 is M and … and X_{47} is M and X_{48} is M and X_{49} is M THEN Label is Male"

After generating the FRBCS model, we can perform prediction over data testing of 5-fold cross validation. Table 3 shows the results over 15 points of data testing on 5 simulations. It should be noted that "Ma" and "Fe" mean male and female, respectively. So, in these experiments the accuracies of 5 simulations are 73.33%, 86.67%, 86.67%, 80%, and 53.33%, and their average is 76%. Moreover, we can see that the system has failed to predict 18 samples. From these incorrect prediction, we can see that it has predicted male for 14 female labels and predicted female for 4 male labels. It means that the system pretend to predict male rather than female. However, in general speaking, we can state that the system can provide reasonable results.

Some comparison can be also done with the some previous works. For example, in the research of [6] that used Gaussian Mixture Models, the results showing the gender detection rate is 67.57%. The research using random forest and kernel discriminant analysis provided the accuracy 74.05% for gender prediction [7]. Moreover, in the research conducted by Hartley [21] the average success rate for classifying gender from handwriting is between 57% and 78%. It

mean that the proposed model can improve the accuracy rate of the previous research. Furthermore, we also provide the understandable model built in fuzzy rule based systems.

Table 3. Results: Predicted and Actual Values on 5-Fold Cross Validation

5-fold cross validation									
1st fold		2nd fold		3rd fold		4th fold		5th fold	
Act.	Pred.	Act.	Pred.	Act.	Pred.	Act.	Pred.	Act.	Pred.
Ma	Fe	Ma	Ma	Fe	Fe	Fe	Ma	Fe	Ma
Ma	Ma	Fe	Fe	Ma	Ma	Fe	Fe	Fe	Fe
Fe	Fe	Fe	Fe	Fe	Fe	Fe	Fe	Fe	Ma
Fe	Fe	Ma	Ma	Ma	Fe	Fe	Fe	Ma	Ma
Ma	Ma	Fe	Fe	Ma	Ma	Fe	Ma	Fe	Ma
Fe	Fe	Fe	Ma	Fe	Fe	Fe	Ma	Ma	Ma
Fe	Fe	Fe	Fe	Ma	Ma	Ma	Ma	Ma	Ma
Fe	Fe	Ma	Ma	Fe	Fe	Ma	Ma	Fe	Ma
Fe	Ma	Ma	Ma	Fe	Fe	Ma	Ma	Fe	Ma
Ma	Ma	Fe	Fe	Fe	Ma	Ma	Ma	Ma	Ma
Ma	Ma	Fe	Ma	Ma	Ma	Fe	Fe	Fe	Fe
Ma	Fe	Ma	Ma	Ma	Ma	Fe	Fe	Fe	Ma
Fe	Fe	Ma	Ma	Ma	Ma	Ma	Ma	Ma	Ma
Ma	Ma	Fe	Fe	Ma	Ma	Ma	Ma	Fe	Fe
Fe	Ma	Ma	Ma	Ma	Ma	Ma	Ma	Ma	Fe

6. Conclusions

The main contributions on this research in on developing a model and its implementation to predict gender from handwriting have been proposed. It consists of the following steps: determining some important features based on graphology techniques, collecting data, image processing on obtained data, and implementing Fuzzy Rule-Based Classification Systems using Chi's Algorithm. Moreover, some experiments are provided to validate and test the model and system by conduction 5-fold cross validation. The simulations show reasonable results, where the average accuracy is 76%.

In the future work, some methods and their software libraries can be utilized to predict the gender according to the handwriting, as the following research: RoughSets [22-23], a combination of Dempster Shafer and Naïve bayes [24], and Fuzzy C-Means [25].

References

[1] Oliveira LS, Justino E, Freitas C, Sabourin R. *The Graphology Applied to Signature Verification*. Proceeding of 12th Conference of the International Graphonomics Society. Salerno, Italy. 2005; 286-290.
[2] Bradley S. Handwriting and Gender: a Multi-use Data Set. *Journal of Statistics Education*. 2015; 23(1): 1-15.
[3] Weisberg YJ, DeYoung CG, Hirsh JB. Gender Differences in Personality Across the Ten Aspects of the Big Five. *Frontiers in psychology*. 2011; 2(178): 1-11.
[4] Crepieux-Jamin J. *L´ecriture et le caractere*. Fourteenth Edition. Presses Univ. de France. 1951.
[5] Hassaine A, Maadeed SA, Aljaam J, Jaoua A. *ICDAR 2013 - Competition on Gender Prediction from Handwriting*. Proceeding of 12th International Conference on Document Analysis and Recognition. Washington, USA. 2013; 1417-1421.
[6] Liwicki M, Schlapbach A, Bunke H. Automatic Gender Detection using On-line and Off-line Information. *Pattern Analysis and Applications.* 2011; 14(1): 87-92.
[7] Al Maadeed S, Hassaine A. Automatic Prediction of Age, Gender, and Nationality in Offline Handwriting. *EURASIP Journal on Image and Video Processing.* 2014*;* 2014(1): 1-10.
[8] Zadeh LA. Fuzzy sets. *Information and control*. 1965; 8(3): 338-353.
[9] Chi Z, Yan H, Pham T. Fuzzy Algorithms: with Applications to Image Processing and Pattern Recognition. World Scientific. 1996; 10.
[10] Pedrycz. W. Fuzzy Modelling: Paradigms and Practice. Kluwer Academic Press. 1996.
[11] Sugeno M, Yasukawa T. A Fuzzy-logic-based Approach to Qualitative Modeling. *IEEE Transactions on Fuzzy Systems*. 1993; 1(1): 7-31.
[12] Wang LX, Mendel JM. Generating Fuzzy Rules by Learning from Examples. *IEEE Transactions on Systems, Man, and Cybernetics*. 1992; 22(6): 1414-1427.

[13] Riza LS, Bergmeir C, Herrera F, Benitez JM. FRBS: Fuzzy rule-based Systems for Classification and Regression in R. *Journal of Statistical Software*. 2015; 65(6): 1-30.
[14] Panjaitan SD, Hartoyo A. A Lighting Control System in Buildings based on Fuzzy Logic. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2013; 9(3): 423-432.
[15] Suardinata S, Bakar KBA. A Fuzzy Logic Classification of Incoming Packet for VoIP. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2010; 8(2): 165-174.
[16] Lubis A J, Susanto E, Sunarya U. Implementation of Maximum Power Point Tracking on Photovoltaic Using Fuzzy Logic Algorithm. *TELKOMNIKA Telecommunication Computing Electronics and Control*. 2015; 13(1): 32-40.
[17] Russ JC, Neal FB. The Image Processing Handbook. Seventh Edition. CRC press. 2016.
[18] Otsu N. A Threshold Selection Method from Gray-level Histograms. *IEEE transactions on systems, man, and cybernetics*. 1979; 9(1): 62-66.
[19] Hou HS. Digital Document Processing. John Wiley & Sons, Inc. 1983.
[20] Arlot S, Celisse A. A Survey of Cross-validation Procedures for Model Selection. *Statistics surveys*. 2010; 4: 40-79.
[21] Hartley J. Sex Differences in Handwriting: A comment on Spear. *British educational research journal*. 1991; 17(2):141-5.
[22] Riza LS, Janusz A, Bergmeir C, Cornelis C, Herrera F, Ślezak D, Benítez J M. Implementing Algorithms of Rough Set Theory and Fuzzy Rough Set Theory in the R package "RoughSets". *Information Sciences*. 2014; 287: 68-89.
[23] Nazir S, Shahzad S, Riza LS. Birthmark-based Software Classification using Rough Sets. *Arabian Journal for Science and Engineering*. 2017; 42(2): 859-871.
[24] Mulyani Y, Rahman EF, Riza LS. *A New Approach on Prediction of Fever Disease by using a Combination of Dempster Shafer and Naïve bayes*. 2016 2nd International Conference on Science in Information Technology (ICSITech). Balikpapan. 2016; 367-371.
[25] Riza LS, Awaludin R, Sutarno H, Munir, Wibawa AP. A Model for Auto Generating Sets of Examination Items in Educational Assessment by using Fuzzy C-means. *World Transactions on Engineering and Technology Education*. 2017; 15(2): 114-119.