ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 5, May 2016**

# Fuzzy clustering based data reduction for improvement in classification

Sayara Bano, Shweta Bandhekar

Department of Computer Science & Technology Rungta Collage of Engineering & Technology

Bhilai, Chhattisgarh, India.

**ABSTRACT:** In this present applying a pre-processing approach for mixed, nominal dataset and continuous dataset. In this paper we are going to propose that the classification technique for classifying the mixed integer, nominal and continuous attributes. For converting the nominal attribute we are applying additionally illustrate method to convert integer and continuous data into nominal attribute. If the data range is small we will directly convert the nominal attribute, if the range is high we will clustered the data. The Fuzzy C-mean clustering algorithm was applied to classify the range of the attribute. By using Fuzzy C-Mean clustering algorithm we will cluster the data of large range of attribute. Ones the data set is pre-processed, we can use a support vector machine (SVM) for classification. We can show the improved accurateness and effectiveness of accessible approach for mixed knowledge in analysis of the classification of the original dataset and nominal dataset.

**KEYWORDS**: Classification, pre-processing data, Fuzzy C-mean clustering algorithm (FCM), Clustering the attribute values

## I. INTRODUCTION

With the rapid development of techniques in information acquisition and storage, spatial knowledge bases store associate degree increasing amount of space-related knowledge. These data, if analysed, can reveal useful patterns and info. The two most often used techniques for info discovery and grouping objects with similar properties are: classification and cluster. Both are typically confused, but some important variations exist between them. Given a relational table, a conventional cluster formula group's tuples, each of that is characterized by a collection of attributes, into clusters are based on resemblance. Spontaneously, and tuples in a cluster are further reasonably like each other than those happiness to wholly totally different clusters. It has been shown that clustering is a very important processing application [1]. On the contrary, classification is a supervised learning strategy that emphasizes on building models able to assign new instances to one of a bunch of well-defined classes. Classification problems have been intensively studied by a varied cluster of researchers along with statisticians, engineers, biologists, computer scientists. There are reasonably methods for determination classification problems Including fuzzy logic, neural networks (NN), and support vector machines (SVM) and principal component analysis (PCA), and linear programming tolerant rough sets etc. The data points in typical processing problems are delineate by a variety of attributes which will have numerical or nominal values [2]. The numerical attributes can have real or separate values, while the separate attributes can be binary or range. This paper focuses on mixed datasets where all varieties of attributes are classified by the classification techniques.

Mixed data are challenges in every supervised and unsupervised learning [3]. Many studies assume concerning distances and similarity to manage mixed information like weak matching distance. Our goal is to propose a pre-processing methodology for the mixed information in multi cluster classification. The data consists of collections of 'Observations' depicted as points in Rd. The mechanism of the vectors gets up for the results. The goal of the data analysis is to achieve logical explanations characteristic observations. It was shown among the literature that within the case of binary knowledge, this goal can be achieved by the use of partially printed scientist functions. The classification based on scientist operate is established to attain success [4].

 DOI: 10.15680/IJIRCCE.2015. 0401001 1111

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 5, May 2016**

Classification algorithms are dynamic into further powerful and wide used attributable to advances in computing hardware and management systems. Some of the classification algorithms use heuristics some use exact ways that, and some use hybrid methods [5]. The classification algorithms are experienced on standard datasets that may exhibit wholly totally different characteristics relying on the applying house. The results are very sensitive to the character and distribution of the info. The attribute of a given dataset can have wholly totally different characteristics. Some examples from the addressed dataset in this paper (see II-C) are:

- •Continuous attribute: such as in the Pima Indian dataset where the 'Diabetes-pedigree-function' attribute has values between 0.078 and 2.42.
- •Nominal attribute: such as in the Servo dataset where the 'motor' attribute can take the following values: A, B, C, D, or E.
- •Integer attribute: such as in the Vehicle dataset where the 'Scaled Radius of Gyration' attribute has values between 109 and 268.

This paper presents pre-processing strategy for mixed datasets. The main effect of feature all nominal information is that the explosion among the variety of choices. In this paper, we proposed a classification of mixed integer and nominal, continuous data for pre-processing before classification. Selection of the classification methodology is together really critical; this methodology have to be compelled to be able to classify instances with a very sizable quantity of choices. We illustrate our approach with well-known SVM methodology. The nominal process shows that the all nominal dataset provides higher accuracy than the original dataset. We conduct an applied arithmetic analysis to appear at the goodness of match of each processed attribute to look at whether or not or not information distribution has any result on classification accuracy with or whereas not pre-processing.
The remainder of this paper is structured as follows: section two discusses the used information classification ways that and platform, section three presents the planned feature binarization methodology, section four presents and discusses the obtained results, and finally section five summarizes the

## II. MATERIALS

There are all classification techniques have advantages and disadvantages, that are a lot of or less vital in line with the information which are being analyzed. SVMs can be a useful gizmo for giant datasets. In this section, we gift the knowledge classification strategy are used in this study: (1) Support Vector Machine classification technique (SVM) and (2) Fuzzy c-mean (FCM) clustering techniques.
The MATLAB 8.5.0.197613 (R2015a) environment was used to perform experiments with the on top of classifiers. It is a group of machine learning algorithms for data processing tasks. It aims to build a state of the skill of facility for developing the techniques of machine learning and investigating their application. . In this study, MATLAB is used to try to the pre–processing data, cluster the attributes and classify the datasets (see section III).

### 2.1. Support Vector Machine (SVM):

The Support Vector Machine (SVM) is a classification technique [6]. SVM construct a system model using a set of given coaching samples for the classification and prediction of the output supported the coaching samples and input samples. In many pattern classification applications, especially with massive datasets, SVMs have exhibited excellent performance [7]. Support vector machines (SVMs) are a set of connected supervised learning strategies used for classification and regression. Viewing input data as 2 sets of vectors in associate degree n–dimensional house, an SVM construct 2 separated hyper planes in that house, one which maximizes the margin between the 2 knowledge sets. The SVM is calculating the margin, and two parallel hyper planes are completed, one on each facet of the separating hyper plane, which are pushed against the knowledge sets.

 DOI: 10.15680/IJIRCCE.2015. 0401001 1112

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 5, May 2016**

In the previous couple of years, there has been a flow of interest in Support Vector Machines (SVMs). SVMs have empirically been shown to offer smart generalization performance on a good form of issues like written character recognition, face detection, pedestrian detection, and text categorization. However, the use of SVMs remains limited to any cluster of researchers. One possible reason is that coaching algorithms for SVMs are slow, especially for massive issues. Another explanation is that SVM coaching algorithms are advanced, subtle, associate degree difficult for an average obtain to implement.

Intuitively, a good separation is achieved by the hyper plane that has the most important distance to the neighbor data-points of each category, since in general the larger the margin the higher the generalization error of the classifier. In many application fields, the SVM has been considered as one of the primary inexpensive strategies for two-class classification issues [8]. It is ready to generate a separating hyper surface so as to maximize the margin and produce smart generalization ability. However, the SVM has two vital drawbacks: a combination of SVMs need to be utilized in order to unravel the multi-class classification drawback, and some approximation algorithms are utilized in order to cut back the procedure time for SVMs whereas learning the big vary of knowledge. To overcome these problems, many variants of SVM have been urged, including the use of SVM ensemble with sacking or boosting instead of employing a single SVM [9].

- For linearly Separable data, SVM finds a separating hyper plane which separates the data with the largest margin.

$$x \in R^1 \rightarrow \emptyset(x) \in R^H$$

**2.2. Fuzzy C-Mean clustering rule:**

Clustering is a method for grouping a group of objects into categories or clusters so the objects at intervals a cluster have high similarity, but are terribly dissimilar to objects in alternative clusters [10]. Various varieties of agglomeration strategies are planned within the literature [11–13]. All of these methods share a typical feature: they're unattended. Because of this, the clustering results would like to be valid. The cluster validation problem involves measurement however well the agglomeration results mirror the structure of the knowledge set, which is associate degree vital issue in clustering analysis. The most important factor of the structure is that the variety of clusters. Since most basic clustering algorithms assume that the quantity of clusters in an exceedingly knowledge set may be a user-defined parameter (one that's tough to line in sensible applications), the common approach is a repetitive trial-and-error process. The trial-and-error process performs the model choice according to the terms used [12]. In fact, the number of clusters may be a parameter associated with the quality of the cluster structure. In this paper, we are interested within the drawback of cluster validation within the context of partitioning-based agglomeration algorithms. In other words, the clustering rule is run with completely different initial values for the variety of clusters and therefore the results are compared so as to work out the foremost acceptable variety of clusters. For this purpose, validity indices have been proposed within the literature [14–18].

In the work reported here, we are notably interested in the fuzzy C-means (FCM) rule. Because of its construct of fuzzy membership, FCM is able to deal a lot of effectively with outliers and to perform membership grading, which is terribly vital in apply. FCM is one of the foremost widely used agglomeration algorithms. Several validity indices have been planned within the literature to be used with the FCM agglomeration rule. Early indices such as the partition coefficient and classification entropy create use solely of membership values and have the advantage of being simple to calculate. Now, it is widely accepted that a more robust definition of a validity index should contemplate each the compactness at intervals every cluster and therefore the separation between clusters. Most existing validity indices are economical in

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 5, May 2016**

police investigation then variety of clusters once the knowledge in numerous clusters doesn't overlap. However, we can see that for overlapping knowledge, their behaviour could be unpredictable. In this paper, we report 2 contributions to cluster analysis.

First, we propose a new rule for agglomeration whereas mechanically determinative the quantity of clusters. The new algorithm improves the standard model choice method by reducing the randomness within the format of cluster centres at the start of every agglomeration section. For this purpose, splitting ways have been designed and combined with the essential agglomeration rule so for every candidate variety of clusters, the clustering method will be administered beginning with the antecedent obtained clusters. Second, we propose a new validity index that's economical even once clusters overlap one another. The new validity index of compactness and separation. We report the take a look at results yielded by the index in a model choice method employing a knowledge set from the general public domain and a number of other generated knowledge sets. These results provide associate degree analysis of the new index beneath the condition of overlapping clusters, an empirical comparison with alternative indices, and an analysis of new rule in terms of numerical stability and time period price.

**Algo1:** Basic FCM algorithm

(1) Randomly select 'c' cluster centers.

(2) calculate the fuzzy membership 'µij' using:

$$\mu_{ij} = 1 / \sum_{k=1}^{c} (d_{ij} / d_{ik})^{(2/m-1)}$$

(3)Compute the fuzzy centers 'vj' using:

$$v_j = (\sum_{i=1}^{n} (\mu_{ij})^m x_i) / (\sum_{i=1}^{n} (\mu_{ij})^m), \forall j = 1, 2, .....c$$

(4) Repeat step 2 and 3 until the minimum 'J' value is achieved or ||U (k+1) - U (k) || < β.

**2.3. DATASETS:**

Our purpose in this proposed method is simply to evaluate the performance of the proposed features on several data sets. To obtain results we need to use a range of data sets that encompass a variety of attributes, range and types, and we hope that the seven sets detailed below provide just such a variety. They are all from UCI Machine Learning Repository [19]. Description of each dataset is given below:

1) Balance scale weight and distance database: This data set was generated to model psychological experimental results. Each example is classified as having the balance scale tip to the right, tip to the left, or be balanced. The attributes are the left weight, the left distance, the right weight, and the right distance. The correct way to find the class is the greater of (left distance* left-weight) and (right-distance * right weight). If they are equal, it is balanced. Number of Instances: 625 (49 balanced, 288 left, 288 right). Number of Attributes: 4.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 5, May 2016**

2) Pima Indians Diabetes: The goal is to diagnose whether the patient shows signs of diabetes according to World Health Organization criteria. All patients here are females at least 21 years old of Pima Indian heritage. Number of Instances: 768 (500 healthy and 268 with diabetes). Number of Attributes: 8.
3) Statlog (Heart): The Statlog heart disease dataset was taken from 270 samples belonging to patients with heart problem while the remaining 150 samples are of healthy persons. Number of Attributes: 13.
4) Servo: The data was from a simulation of a servo system involving a servo amplifier, a motor, a lead screw/nut, and a sliding carriage of some sort. Number of Instances: 167. Number of Attributes: 4.
5) Lenses: The lenses data set has four variables and three classes. The variables are age of the patient, spectacle prescription, and astigmatic and tear production rate. Classes which need to be predicted are hard contact lenses, soft contact lenses and no contact lenses. The data consists of 24 samples so that it can be considered quite small.
6) Breast Cancer Wisconsin (Original): Original Wisconsin Breast Cancer Database. Number of Instances: 699 (Benign: 458, Malignant: 241). Number of Attributes: 10.

## III. METHODOLOGY

This section presents the proposed pre-processing techniques for mixed dataset. An illustrative example can show however the methodology treats the dataset. In classification problems, we will realize varied forms of knowledge sets; each one will have an outsized, moderate, or small range of attributes. The attributes can have categorical, integer, or continuous values that can be discovered during a dataset. The nature of the attribute and therefore the distribution of its values rely upon the domain of the dataset. In this paper, we propose a pre-processing method that can prepare the datasets to perform the classification task. The simplest non binary attributes are the alleged 'nominal' ones. A typical nominal attribute is' male, whose values can be 'gender', and so on

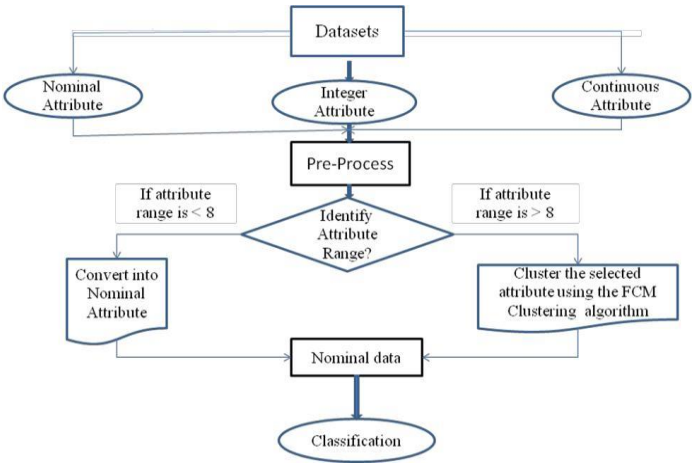


**Figure:** Flowchart of proposed approach for classification with pre-processing.

For a selected dataset, each categorical or number attribute are regenerate into a nominal attribute. For a categorical attribute, we already understand all its attainable values. Thus, it is easy to convert it into a nominal attribute. For example, a colour attribute which will takes three totally different values: 'blue', 'gray' or 'yellow' is converted into 3

 DOI: 10.15680/IJIRCCE.2015. 0401001 1115

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 5, May 2016**

nominal attributes. The first attribute takes the worth '1' if the colour is 'blue', '0' otherwise. The second attribute takes the value '1' if the colour is 'gray', '0' otherwise. And the third attribute takes the worth '1' if the colour is 'yellow', '0' otherwise. However, a number attribute will take values in a little or massive vary. Thus, it is difficult to convert all its attainable values into nominal values if it's an outsized vary. If the integer attribute has a small range, it is considered a categorical attribute; and it follows identical steps as a categorical attribute. If the integer attribute has a large range, we classify its values into k categories. In that case, according to the attribute distribution, it takes the value once we contemplate this attribute as a categorical attribute and it's simple to convert it into a nominal attribute. We use the FCM rule to confirm k classes; k is decided by the FCM rule within the bunch step of associate degree attribute with an outsized vary. The value of k depends on the distribution of the number attribute values. Figure 1 summarizes all the steps of the approach.

### 3.1. Illustrative example

In this example, we are considering an artificial dataset that contains mixed integer, nominal and continuous attributes as shown in Table 1. There are ten instances and ten attributes. This data is composed by six number attributes (A, D, F, H, I and J), 2 continuous attribute (B and G) and two nominal attribute (C and E). We initial cluster this knowledge discrimination the well well-known cluster methodology FCM and show the cluster lead to the last column of table one. Here, we have an appropriate vary for the A, F, and J attributes (A: from 1 to 6), (F: from 1 to 6 and J: from zero to 6). Thus, it can be simply regenerate into nominal attribute and therefore the D, H and I attributes have an outsized range (D: 19 to 25; H: 14 to 45 and I: from 0 to 786). Thus we have a tendency to have to cluster its values and that we get the nominal attributes D, H and I. The values of the nominal attributes C and E are well proverbial, and they may be converted into nominal attributes. The D and G attributes can be used as continuous attributes.

Table 2 shows the Nominal Dataset when the initial step of the pre-processing has done. The D attribute presents two totally different values (cluster1 and cluster2). These values are obtained when applying the FCM clustering algorithm on the D attribute that divide it into two clusters. The H attribute presents also 2 totally different values (cluster1 and cluster2). These values are obtained when applying the FCM clustering algorithm on the H attribute. The I attribute presents two totally different values (cluster1 and cluster2). These values are obtained when applying the FCM clustering algorithm on I attribute. The classification accuracy of the presented sampleon SVM is given in table 3. The SVM is used for classifying the mixed datasets.

## IV. RESULT

The performance evaluation of the feature FCM method that we have a tendency to propose is given in Table four. The left part of this table shows the used datasets (name and range of classes). The middle part shows the quantity of every sort of attribute (nominal, continuous and integer). And the right a part of table 4 shows the typical of the properly classified information (original dataset, nominal dataset and binary dataset). All the used datasets are variable. Instances with missing attributes values were removed. In the initiative, we classify the original dataset and gift the accuracy within the 1st column of the correct a part of Table four.

In the second step, the integer attributes with little vary values were reworked to nominal attributes, and the integer attributes with giant vary values were classified with the FCM algorithmic rule. These attributes are as follows: (i) All attributes in (i) 'Number Of Times', 'Plasma Glucose' 'Diastolic', 'Triceps', '2-HourSerum', 'Age' in Pima Indian dataset, (iii) 'age', 'resting blood pressure', 'serum cholesterol in mg/dl' and 'Maximum Heart rate achieved' in Heart

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 5, May 2016**

dataset, (iv) and no attribute was classified with the FCM in the Balance scale, Servo, Lenses dataset. At this stage, we have a nominal dataset. The average of the correctly classified information is equal or superior to the classification of the first dataset. Finally, in the third step, we have all nominal attributes.

In the last column of table 4, we gift the improvement within the classification accuracy with the nominal dataset. We observe 2 distinct characteristics in the results as shown in Table four. The first class of results includes Balance scale, Servo, Lens datasets. In these datasets, there is no accuracy improvement within the classification of the nominal data. We will observe that those datasets were subject to modification of distinct attributes into nominal attributes (without doing the bunch of the attributes). After applying the FCM method, the classification accuracy improves. That means that the FCM has an improved the result of datasets. The second category contains: cancer, Pima Indian and Heart. In these datasets, we realize distinct attributes with giant ranges and that we cluster them with the FCM algorithmic rule. As a result, the attributes were transformed to nominal. The classification accuracy of the nominal dataset is much on top of the first dataset. Yet, the classification accuracy of the nominal dataset is even better. We analyze the distribution of every clustered attribute of those datasets.

**Table 3:** Classification accuracy of the presented sample

| **Data pre-processing** | **Classification Accuracy (%)** |
|---|---|
| Original dataset | 74 |
| Nominal Dataset | 81.2 |

**Table 4**: Numerical result on benchmark dataset

| Dataset | #of attribute | | | Correctly classified (%) | |
|---|---|---|---|---|---|
| Data name | Nom. | Cont. | Int. | Original value | Nom. Value |
| Balance Scale | 0 | 0 | 4 | 97 | 97 |
| Servo | 2 | 0 | 2 | 18 | 18 |
| Lenses | 0 | 0 | 4 | 90 | 90 |
| Cancer | 0 | 0 | 9 | 89.3 | 94.3 |
| Pima Indian | 0 | 2 | 6 | 75 | 96 |
| Heart | 0 | 1 | 12 | 80 | 92 |

## V. CONCLUSION

In this paper, we are presenting the result of pre-processing method for the classification accuracy. Preprocessing is applied for all integers, nominal and continuous dataset. These data points are transferred into nominal attribute. Pre-processing methodology conferred in section 3. We applied the pre-processing approach to six type of dataset in multi group classification drawback with mixed attribute. There all instances with missing were removed. The integer attribute with small range values were transferred to nominal attributes. Integer attribute with a large range of values were classified by the FCM clustering method. At this stage, we get a nominal dataset. Six datasets were used to test the pre-processing method. The results show that the nominal dataset provides better accuracy than the original dataset.

It is observed that majority of the information exhibit separate Uniform distribution, representation of the knowledge as original doesn't improve the classification accuracy. There may be some improvement within the classification

 DOI: 10.15680/IJIRCCE.2015. 0401001 1117

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

*(A High Impact Factor, Monthly, Peer Reviewed Journal)*

**Vol. 4, Issue 5, May 2016**

accuracy if it is improved by nominal data, procedure presented in this paper. When the knowledge incorporate a huge variation within the distribution profile (such as cancer, Pima Indian and Heart) the pre-processing of data to convert into nominal attribute for improvement the classification accuracy.

## REFERENCES

1. WH Au, KCC Chan, et al. The attribute clustering for assemblage, assortment, and classification of gene appearance data, Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2005; 2: 83–101.
2. F Uney, M Turkay, A mixed-integer series method to multi-class data classification problem, European Journal of Operational Research, 2006; 173: 910 – 920.
3. E Tuva, GC Runger, The scoring levels of the categorical variables with heterogeneous data, Intelligent Systems IEEE, 2004; 19: 14–19
4. E Boros, P Hammer,et al. The logical analysis of the numerical data, Mathematical Programming, 1997; 79; 63–190.
5. H Said, C Habib, et al. A mathematical programming based on the procedure for breast cancer classification, Journal of Mathematical Modelling and Algorithms, 2010; 9: 247–255.
6. C Cortes, V Vapnik, The support-vector networks, Machine (SVM) Learning, 1995; 20: 273–297.
7. J Terzic, C Nagarajah, et al. The fluid level of measurement in dynamic environments using a single ultrasonic sensor and support vector machine (SVM), Sensors and Actuators A: Physical, 2010; 161: 278 – 287.
8. Y Yajima, A linear programming approaches for multicategory support vector machines (SVM), European Journal of Operational Research, 2005; 162: 514 – 531.
9. HC Kim, S Pang, et al. The constructing support vector machine (SVM) ensemble, Pattern Recognition, 2003; 36: 2757 – 2767.
10. J Han, M Kamber, et al. Datamining: The concepts and Techniques, Morgan Kaufmann Publishers, San Francisco, 2001.
11. F Hoppner, Fuzzy Cluster Analysis: Methods for Classi5cation, Data Analysis and Image Recognition, Wiley, New York, 1999.
12. AK Jain, RC Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cli;s, NJ, 1988.
13. R Krishnapuram, O Nasraouui, et al. The fuzzy C-spherical shells algorithm is used for cluster the data: a new approach, IEEE Trans.Neural Networks, 1992; 3: 663–671.
14. X Xie, G Beni, A authority measure for fuzzy clustering, IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) 1991; 13: 841–847.
15. Y Fukuyama, M Sugeno, A new techniques of choosing the number of clusters for the fuzzy C-means (FCM) method, in: Proceedings of Fifth Fuzzy Systems Symposium, 1989; 247–250.
16. M Rezae, B Letlieveldt, et al. A new cluster strength index for the fuzzy c-means, Pattern Recogn. Lett, 1998; 19: 237–246.
17. H Rhee, K Oh, A validity measure for fuzzy clustering (FCM) and its use in selecting optimal number of clusters, in: Proc. IEEE, 1996; 1020–1025.
18. Baraldi, P Blonda, A survey of the fuzzy clustering (FCM) algorithms for pattern recognition—part I, IEEE Trans. Syst. Man, Cybern. SMC-29 1999; 778–785.
19. Frank, A Asuncion, an UCI machine learning repository, 2010.

 DOI: 10.15680/IJIRCCE.2015. 0401001 1118