

Feature Evaluation Based Fuzzy C-Mean Classification

Mostafa A. Salama
Department of Computer Science
British University in Egypt
Cairo, Egypt
E-mail: mostafa.salama@gmail.com

Aboul Ella Hassanien and Aly A. Fahmy
Faculty of Computers and Information
Computer Engineering
Cairo University
E-mail: aboitcairo, aly.fahmy@gmail.com

Abstract—Fuzzy C-Means Clustering, FCM, is an iterative algorithm whose aim is to find the center or centroid of data clusters that minimize an assigned dissimilarity function. The degree of being in a certain cluster can be defined in terms of the distance to the cluster-centroid. The domain knowledge is used to formulate an appropriate measure. However the Euclidean distance is considered as a general measure for such value. The calculation of the Euclidean distance doesn't take into consideration the degree of relevance of each feature to the classification model. In this paper, scoring methods like ChiMerge and Mutual information are used in the FCM model to improve the calculation of the Euclidean distance. Experimental results demonstrate the better performances of the improved FCM on UCI benchmark data sets rather than the ordinary FCM, where the ordinary FCM uses in classification either all features or the most important features while the improved FCM uses all the features but the Euclidean Distance will be calculated according to the relevance degree of each feature.

Index Terms—Fuzzy C-Means; Feature selection; ChiMerge.

I. INTRODUCTION

Fuzzy C-Mean (FCM) is a model of data clustering which allows one piece of data to belong to two or more clusters. It is based on minimization of an objective or dissimilarity function [1], [2]. With Fuzzy C-Means, the centroid of a cluster is computed as being the mean of all points, weighted by their degree of belonging to the cluster. The degree of being in a certain cluster is related to the inverse of the distance to the cluster. For high dimensional data, a popular measure is used for calculating the distance which is the Minkowski metric [3], Euclidean distance is a special case of such equation. The calculation of the Euclidean distance ignore the degree of relevance of each feature to the classification problem and treats all the features equally. On the other hand, high number of features may contain irrelevant or redundant features to the classification [4]. To handle this curse of dimensionality problem, feature reductions methods are used. The number of features are reduced to the minimum value either by selecting the best features or extracting a lower number of feature from the higher ones. Consequently, applying the classifier on the resulted and reduced features should show a better performance. Classifiers could use the ChiMerge and the Mutual Information (*MI*) methods as feature selection methods in the pre-processing step of the Classification system design procedure [5]. Chi-square testing

χ^2 is a well-known discrete data hypothesis testing method from statistics, for continuous data sets, the ChiMerge method can automatically select a proper χ^2 values and determining the intervals of numeric attributes [6]. On the other hand *MI* indicates the strong informative relationships among the attributes. In this paper, the ChiMerge or *MI* methods are applied to evaluate the worth of each attribute by computing the value of the chi-squared statistic with respect to the class. The resulted value is scaled to the range or interval between 0 and 1 and will be defined as ranks. Previously, the result from the feature selection methods are either 0 or 1 and used in the classifier, but here we put into account all features but with different ranks. Then these ranks are used in the calculation of the Euclidean distance used in the Fuzzy C-Mean model. The new factor is that the features with less chi-square χ^2 or *MI* values will have a less influence on the classification procedure rather than being removed completely and rather than using the total set of feature. According to the results this modification generates a better classification performance. Also the results in the case of using ChiMerge and *MI* methods are compared and it shows that ChiMerge has a better results.

The rest of this paper is organized as follows. Section II discusses the related work. Section III explains the Fuzzy C-Mean clustering model. Section IV explain the calculation of the distance to the cluster center. Section V shows the feature evaluation method used in details which are the ChiMerge and Mutual information methods. section VI shows how the chi-Square values will be used in the Fuzzy C-Mean classification. Results and comparisons with other classifiers are illustrated in section VII. Conclusion and future work is discussed in section VIII.

II. RELATED WORK

Most of the improvements in the Fuzzy C-Mean models are applied on the selection the most relevant or informative features. The benefits of feature selection [7] includes a lot of aspects like facilitating data visualization and data understanding, reducing the measurement and storage requirements, providing faster and more cost-effective classifiers, and defying the curse of dimensionality to

improve prediction performance. A paper like in [8] has an addition is that the fuzzy feature clustering and feature selection are based on the interdependence among features using chi-square methods. Other feature selection methods like in [9] has improved the *MI* formula for feature ranking which is used for feature selection in the pattern classification task. The problem of the current work in this area is that it stresses on the improvement of the feature selection methods, while ignores the use of the feature ranks of the selected features in the classification models. A similar study has been applied in [10] that propose an approach to feature-weight learning which is based on the gradient-descent technique. This method suffers from the high complexity due to the feature-weight learning which depends on the convergence of iterations in the traditional convergence technique. Also this method has a pre-assumption that the distribution of data is uniform which is not the case in real life data-sets.

III. FUZZY C-MEAN CLUSTERING

The discreteness of each cluster makes the C-Means analytically and algorithmically intractable. Partitioning the data set in a fuzzy manner avoid this problem. The FCM clustering treats each cluster as a fuzzy set and each feature vector is assigned to multiple clusters with some degree of certainty measured by the membership function [11]. The FCM optimizes the following objective function

$$J^m = \sum_{j:1..K} \sum_{i:1..N} \mu_{ji}^m * d_{ij} \quad (1)$$

Where: K is the number of clusters,

N is the number of objects,

m is the fuzziness controller which reduce the influence of small membership values, m is greater than 1,

d_{ij} represents the Euclidean distance between the object i of value x_i and the centroid of cluster j of value c_j which is represented by the following equation:

$$d_{ij} = \|x_i - c_j\| \quad (2)$$

and μ_{ji} denotes the membership of object i in cluster j ; the values of μ_{ji} are initialized randomly such that the following condition is valid:

$$\sum_{j:1..K} \mu_{ji}^m = 1, i = 1, \dots, N \quad (3)$$

and c_j is the centroid of cluster j which it is calculated as follows:

$$c_j = \frac{\sum_{i:1..N} \mu_{ji}^m * x_i}{\sum_{i:1..N} \mu_{ji}^m} \quad (4)$$

After calculating the objective function J^m by equation (1), the value of μ_{ji} is updated by the following equation

$$\mu_{ji}^m = \frac{1}{\sum_{j:1..K} \frac{d_{ij}}{d_{ik}}} \quad (5)$$

The calculation of the J^m is repeated iteratively until the difference between the current J^m and the previous one is smaller than a certain threshold value ε such that:

$$J_{current}^m - J_{last}^m < \varepsilon \quad (6)$$

The correctness of equation (6) indicates the conversion of the objective function.

IV. DISTANCE CALCULATION TO CLUSTERS CENTER

A popular measure used for calculating the distance between any object and the cluster center, is the Minkowski metric. This measure is required for high dimensional data and it is calculated as follows:

$$d_p(x_i, x_j) = \left(\sum_{k=1..D} |x_{ik} - x_{jk}|^p \right)^{\frac{1}{p}} \quad (7)$$

Where x_i and x_j are two objects in the input data set. D is the dimensionality of the input data set, i.e. number of attributes. p is a parameter that represents the type of Minkowski measure. The Euclidean distance [12] is a special case of this measure, where in this case, $p = 1$, while Manhattan metric is the case where $p = 2$. However, there are no general theoretical guidelines for selecting a measure for any given application.

V. FEATURE EVALUATION METHODS

The selected feature evaluation techniques used in this paper are ChiMerge and Mutual Information approaches, which are considered as the most reliable methods for feature selection purpose with no significant difference between them [13].

A. ChiMerge feature evaluation

One of the most popular feature selection methods is the chi-Square χ^2 method, which measures the lack of independence between each attribute A and the target class c . ChiMerge or Chi2-Square is a χ^2 -based discretization method [14]. It uses the χ^2 statistic to discretize numeric attributes repeatedly until some inconsistencies are found in the data that achieves feature selection via discretization. The χ^2 value is calculated for each continuous attribute as follows: Initially, each distinct value of a numeric attribute A is considered to be one interval. The values, intervals, of attribute A are sorted and the χ^2 is applied for every pair of adjacent intervals as follows:

$$\chi^2 = \sum_{i:1..2} \sum_{j:1..k} \frac{(A_{ij} - E_{ij})^2}{E_{ij}} \quad (8)$$

Where:

A_{ij} is the number of values in the i th interval and j th class,
 R_{ij} is the number of values in the j th class = $\sum_{i:1..k} A_{ij}$,
 C_{ij} is the number of values in the i th interval = $\sum_{j:1..2} A_{ij}$,
 N is the total number of values = $\sum_{i:1..2} R_{ij}$,
and E_{ij} is the expected frequency of $A_{ij} = \frac{R_{ij} * C_{ij}}{N}$

Adjacent intervals with the least χ^2 values are merged together, because low χ^2 values for a pair indicates similar

class distributions. This merging process proceeds recursively until all χ^2 values of all pairs exceeds a parameter *signlevel* (initially 0.5). Then repeat the previous steps with a decreasing *signlevel* until an inconsistency rate is exceeded, where two patterns are the same but classified into different categories.

B. Mutual Information feature evaluation

The mutual information *MI* (also called cross entropy or information gain) is a widely used information theoretic measure for the stochastic dependency of discrete random variables. The mutual information $I(A; C)$ between values of feature *A* and the set of classes *C* can be considered as a measure of the amount of knowledge on *C* provided by *A* (or conversely on the amount of knowledge on *A* provided by *C*). In other words $I(A; C)$ measures the interdependence between *A* and *C* where it can be computed as follows:

$$I(A; C) = H(C) - H(C|A) \quad (9)$$

The entropy $H(C)$ measures the degree of uncertainty entailed by the set of classes *C*, and can be computed as

$$H(C) = - \sum_{c \in C} p(c) \log p(c) \quad (10)$$

where $p(c)$ is the probability density function (*PDF*) of *c*. The conditional entropy $H(C|X)$ measures the degree of uncertainty entailed by the set of classes *C* given the set of feature values *X*, and can be computed as

$$H(C|A) = \sum_{c \in C} \int_{a \in A} p(a, c) \log \frac{p(a|c)p(c)}{p(a)} dx \quad (11)$$

The integration in the expression above signifies that the feature space is continuous.

VI. CLASSIFICATION USING THE IMPROVED FUZZY C-MEAN MODEL

By iteratively updating the cluster centers and the membership grades for each data point, FCM iteratively moves the cluster centers to the right location within a data set. But increasing the number of features may lead to ambiguity in training so that it would not even converge. The usual solution for this problem is to eliminate the features using the feature reduction methods, like applying feature selection of the most important and relevant features. After feature selection is applied, the data will be ready for training and testing by the classifier selected. The problem of this procedure is that not all features, even the selected features, are as important as each other to the classifier. On the other hand, in the calculations of the distance each object to the centroid in the Fuzzy C-Mean classifier model, all the attributes are used in the calculation in the same way.

In order to solve this problem, this paper uses a ranking procedure like the chi2-square method to give a percentage of importance to each feature. The results from the chiMerge method, χ^2 values, are processed such that these values range from 0 to 1 and the sum of all values are 1. Then uses the

percentage that corresponds to each feature in the calculation of the Euclidean distance. Such that this percentage will be multiplied by the difference between the values of each object and the centroid corresponding to this attribute. So equation (7) will be adjusted to be as follows:

$$d_p(x_i, c_j) = \left(\sum_{k=1..D} r_k * |x_{ik} - c_{jk}|^p \right)^{\frac{1}{p}} \quad (12)$$

Where r_k represents the rank of attribute *k*, x_i represents the object *i* and c_j represents the centroid of class *j*.

The calculation of the rank values r_k will be as follows:

First, the values that are resulted from the feature selection methods are saved, in our paper these values are the χ^2 values. Then, the degree of importance of each feature is calculated according to the sorting of χ^2 values in an increasing order. The formula of the degree of importance will be in the form:

$$d_k = \frac{D - o_k}{\sum_{k=1..D} o_k} \quad (13)$$

Where o_k represents the order of the attribute from the chi-square method, for example, if $o_k=1$, this means that the attribute *k* is the best attribute, while if $o_k=D$, it means that attribute *k* is the worst attribute.

Then the χ^2 value of each feature *k* multiplied by d_k is divided by the sum of χ^2 values of all feature, and the result is assigned to the rank value r_k . So, r_k will be calculated as shown in the following formula (14):

$$r_k = \frac{\chi_k^2 * d_k}{\sum_{k=1..D} \chi_k^2} \quad (14)$$

In the case when the data set produces all χ^2 values have zero value, so formula (14) should be replaced by following formula (13). As shown in figure (1), the input from the training data set is evaluated by the chiMerge method then the χ^2 values are adjusted as in formula (14), and this input is introduced to Fuzzy C-Mean model for training. The evaluation from the chi2-square method will be used in the calculation of the distance to the centroid. When the termination condition is reached; either by convergence or elapsed number of iterations, the calculated centroids and the adjusted χ^2 values will be used in the testing of the classifier.

This improvement could be applied using the Mutual information method, such that the information gain resulted in equation (9) or the *MI* value for each attribute will be used in formula (14) as follows:

$$r_k = \frac{MI_k * d_k}{\sum_{k=1..D} MI_k} \quad (15)$$

where MI_k value represents the mutual information between feature *k* and the set of target classes. The complete steps of the enhanced FCM model is shown in figure (1), in the steps of the model, scaling of the input data is important as it showed a better results in case of the scaled data than unscaled one.

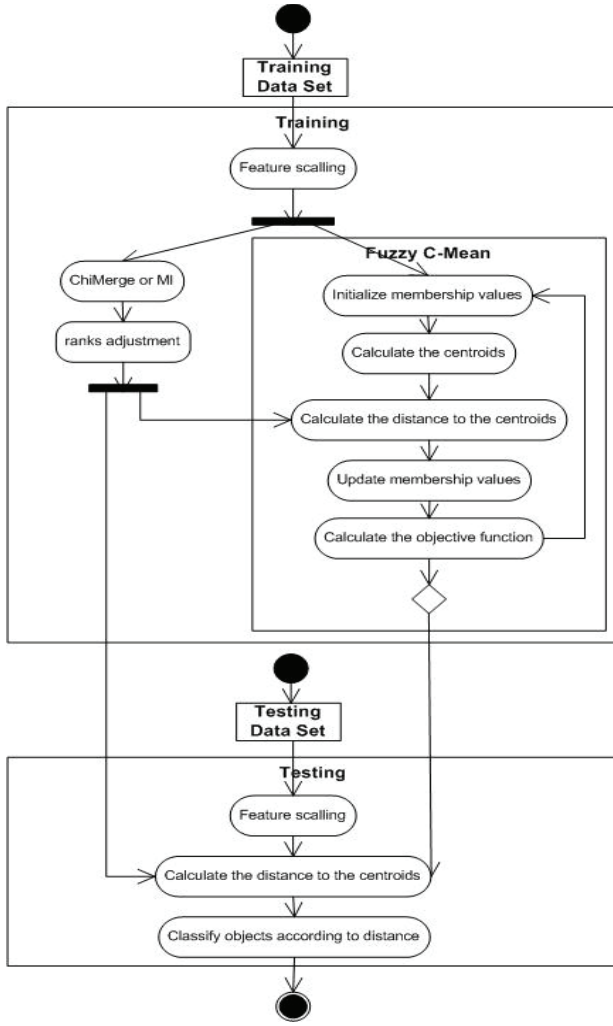


Fig. 1. Improved Fuzzy C-Mean model

VII. EXPERIMENTAL RESULTS AND DISCUSSIONS

The proposed improvement is applied on two data sets and compared to the case when all features are used. Also it is compared to the case when only the important features are used in classification, in order to perform this, the test is applied on the first important feature only, then the first two important features, and so on until all feature are used in classification test. On the other hand the FCM with the proposed improvement is compared to other classifiers.

A. Data sets used in classification

This proposed enhancement is applied on two different data sets to ensure its capability and competency among other classifiers. These data sets are obtained from UCI machine learning repository [15]. Which are the pima-indians-diabetes data set of 536 objects and 8 features, and the Yeast data set of 600 objects and 8 features. 90% of the input data used for training while the rest of 10% is used for testing.

B. Classifiers used in the comparison analysis

The model is tested against other different classification techniques like Bayesian network, Support Vector Machine, Multilayered Neural Network and Decision table.

C. Classification results

1) *Performance Measure*: The classification effectiveness of the improved model is calculated in terms of *precision* and *recall* [16] which are defined as follows:

$$precision = \frac{a}{a+b}, a+b > 0 \quad (16)$$

$$recall = \frac{a}{a+c}, a+c > 0 \quad (17)$$

Where a counts the assigned and correct cases, b counts the assigned and incorrect cases, c counts the not assigned but incorrect cases and d counts the not assigned and correct cases. To evaluate the performance across categories, *F-measure* is averaged. There are two kinds of averaged values, namely, micro average and macro average. The micro value is the average of the two values, *precision* and *recall*, of all classes, while the macro average is the average of all micro values, this average represents the *F-measure*. Table (I) shows the Macro averaged *F-measure* of the improved model when applying the ranking values in the improvement shown in equation (10). The results in the case of pima-indians-diabetes and yeast data sets are compared to the FCM model without the improvement.

TABLE I
THE *F-measure* RESULTS COMPARISON

Data Set Used	IndianDiabetes	Yeast
<i>F-measure</i> of FCM	0.796	0.601
<i>F-measure</i> of Improved FCM Ranking:ChiMerge	0.833	0.642
<i>F-measure</i> of Improved FCM Ranking:MI	0.833	0.627

2) *Applying feature selection before classification*: The following table (II) shows the results when perform the classification using the Fuzzy C-Mean after feature selection. The first row shows the result when only the first feature is used in classification, then the second row shows the results when the first two features are selected, and so on until all the features are used. The last row shows the results appear from the improved FCM model shown in table (I).

3) *Comparison to the other classifiers*: Table (III) shows the results of other classifiers in comparison to the enhanced FCM model where it appears to have best results.

D. Results analysis

Pima-indians-diabetes is continuous multi-variate features data set. So Fuzzy C-Mean model shows a better results

TABLE II
FCM CLASSIFICATION RESULTS OF THE PIMA-INDIANS-DIABETES (PID)
DATA SET AND THE YEAST DATA SET

features used	$F - Measure$ of PID	$F - Measure$ of yeast
1 feature selected	0.740	0.500
2 features selected	0.777	0.600
3 features selected	0.777	0.600
4 features selected	0.777	0.533
5 features selected	0.796	0.533
6 features selected	0.796	0.533
7 features selected	0.796	0.516
8 features selected	0.777	0.566
Ranking features Improved FCM	0.833	0.642

TABLE III
COMPARISON BETWEEN DIFFERENT CLASSIFICATION MODELS
ACCORDING TO ERROR PERCENTAGES AND RUNNING TIME IN SECONDS
AND THE IMPROVED FCM MODEL (I-FCM)

Classifier	pima-indians-diabetes Classification $F - Measure$
I-FCM	0.814
Multilayer Perceptron	0.776
IB1	0.692
BFTree	0.759
Decision Table	0.763
Naïve Bayes	0.746
Bayes Network	0.761
LibSVM	0.527

of 0.777 than other classifiers. As Decision trees and Decision tables tend to perform better when dealing with discrete/categorical features, while Bayes belief models like Bayesian Network (BN) and Naïve Bayesian Network due to the "independence assumption". Moreover, if feature selection is applied before the Fuzzy C-Mean classifier, the maximum classification performance appeared is 0.796 for five selected features. When applying the improvement on the Fuzzy C-Mean model, the classification results is raised more to 0.833. Also the enhancement in the classification performance from the Fuzzy C-Mean to the improved Fuzzy C-Mean model appears to be raised from 0.60 to 0.642 when being applied on the Yeast data set. Chi-Merge ranks leads to a better classification performance than that of Mutual information method due to the better discretization algorithm embedded in the Chi-Merge method. This proves that not all of the selected features have the same importance as each other. In the case of pima-indians-diabetes and yeast data sets, the chi-square method produces different values for each attributes and only few attributes have zero values, formula (14) will be applied without changes. In cases where all features are ordered with different importance but all have the zero rank

values, r_k values could be replaced by d_k values.

VIII. CONCLUSIONS

This paper introduces a new application of the feature selection method in the classification model. It uses the ranks of each feature resulted from feature evaluation methods like ChiMerge and Mutual Information in the calculations of the Euclidian distance as a main parameter in the Fuzzy C-Mean model. This introduced modification in the Fuzzy C-Mean model enhanced the performance of the Fuzzy C-Mean model. The modified FCM has been applied on real life data sets including pima-indians-diabetes and yeast data sets, and it provides the best classification accuracy. The reason of these results is that even the selected features from the feature reduction methods have no equal relevance or contribution to the classifier, so if the degree of relevance in each classifier is taken into consideration, a better results can appear.

REFERENCES

- [1] Tina Gewenigera, Dietlind Zlkebe, Barabara Hammerc and Thomas Villmann, "Median fuzzy c-means for clustering dissimilarity data", Neurocomputing, vol. 73, issue. 7-9, pp. 1109-1116, Mar. 2010.
- [2] Abdelwaddood Moh'd, "Chi Square Feature Extraction Based Svms Arabic Language Text Categorization System", Journal of Computer Science, ISSN 1549-3636, vol. 3, pp. 430-435, 2007.
- [3] Tai-hoon Kim, "Procedure of Partitioning Data Into Number of Data Sets or Data Group", Communications in Computer and Information Science, vol. 78, pp. 104-115, 2010.
- [4] A.G.K. Janeczek, W.N. Gansterer, M. Demel and G.F. Ecker, "On the relationship between feature selection and classification accuracy", JMLR Workshop and Conference Proceedings, vol. 4, pp. 90-105, 2008.
- [5] Thabtah F., Eljinini M., Zamzeer M. and Hadi W., "Naïve Bayesian based on Chi Square to Categorize Arabic Data", In proceedings of The 11th International Business Information Management Association Conference (IBIMA) Conference on Innovation and Knowledge Management in Twin Track Economies, Cairo, Egypt, pp. 930-935, Jan. 4-6, 2009.
- [6] Surendra Kumar, C.S.P. Rao, "Application of ant colony, genetic algorithm and data mining-based techniques for scheduling", Robotics and Computer-Integrated Manufacturing, vol. 25, issue 6, pp. 901-908, Dec. 2009.
- [7] Mostafa A. Salama, Aboul Ella Hassanien and Aly A. Fahmy, "Uni-Class Pattern-based Classification Model", In proceeding of the 10th IEEE International Conference on Intelligent Systems Design and Applications, Cairo, Egypt, Dec. 2010.
- [8] Elham Chitsaz, Mohammad Taheri, Seraj D. Katebi, Mansour Zolghadri Jahromi, "An Improved Fuzzy Feature Clustering and Selection based on Chi-Squared-Test", In proceedings of the International MultiConference of Engineers and Computer Scientists 2009, Hong Kong, vol. 1, Mar. 18-20, 2009.
- [9] Jin-Jie Huang, Yun-Ze Cai, Xiao-Ming Xu, "A parameterless feature ranking algorithm based on MI", Neurocomputing, vol. 71, pp. 1656-1668, 2008.
- [10] Xizhao Wang, Yadong Wang and Lijuan Wang, "Improving fuzzy c-means clustering based on feature-weight learning", Pattern Recognition Letters, vol. 25, issue. 10, pp. 1123-1132, jul. 2004. Neural networks, vol. 23, issue 1, pp. 89-107, Jan. 2010.
- [11] K-L Du, "Clustering: A neural network approach", Neural networks, vol. 23, issue 1, pp. 89-107, Jan. 2010.
- [12] Jing Lia and Bao-Liang Lu, "An adaptive image Euclidean distance", Pattern Recognition, vol. 42, issue 3, pp. 349-357, Mar. 2009.
- [13] J. R. Mndez, B. Corzo, D. Glez-Pea, F. Fdez-Riverola and F. Daz, "Analyzing the Performance of Spam Filtering Methods When Dimensionality of Input Vector Changes", In proceedings of the 5th international conference on Machine Learning and Data Mining in Pattern Recognition, vol. 4571, pp. 364-378, 2007.

- [14] Huan Liu, Rudy Setiono, "Chi2: Feature Selection and Discretization of Numeric Attributes", In proceedings of the Seventh International Conference on Tools with Artificial Intelligence, Virginia, USA, pp.388, Nov. 8, 1995.
- [15] UCI Machine Learning Repository,
<http://archive.ics.uci.edu/ml/datasets.html>.
- [16] Yiming Yang, "An evaluation of statistical approaches to text categorization", Inform Retrieval, vol. 1, pp. 69-90, 1999.