ISBN 978-83-60810-51-4

Unsupervised Partitioning of Numerical Attributes Using Fuzzy Sets

Bogdan Popescu*, Andreea Popescu†, Marius Brezovan‡ and Eugen Ganea§
Faculty of Automation, Computers and Electronics
University of Craiova,
Bd. Decebal 107, Craiova Romania
* Email: bogdan.popescu@itsix.com
† Email: andreea.popescu@itsix.com
‡ Email: brezovan_marius@software.ucv.ro
§ Email: ganea_eugen@software.ucv.ro

***Abstract*—The current paper presents an enhanced partitioning mechanism for numerical data. The efficiency of our method will be illustrated through a solid set of tests that have been performed. We have planned this partitioning phase as an initial step in a more complex algorithm to be further studied and implemented.**

The final goal is to use it for future decision making in automatic image annotation. Fuzzy Sets theory has been used as a base for our clustering algorithm and partitioning. We included this mechanism as a component of a framework we developed for image processing, more exactly for the image segmentation evaluation model we are building.

I. Introduction

One OF the most important problems in data mining is taking decisions based on association rules. In order to achieve this, an efficient partitioning mechanism needs to be used.

This choice can have consistent influence over the final result and we considered this a crucial approach for our further study.

The core topic considered by our paper, partitioning quantitative attributes, has been initially studied by [2]. The author proposed in [2] an initial partitioning into small intervals and combine adjacent intervals into bigger ones so that the domain support to be more relevant, leading to boolean logic when replacements are done for the initial attribute with attribute-interval.

It seems that current associations algorithms introduce further more other problems. The might ignore the elements near the boundary or use not very intuitive for human perception approaches like shape boundary interval.

We proposed through our work an efficient mechanism of partitioning using fuzzy sets logic. We have worked to improve the clustering algorithm starting from the basic Fuzzy C-Means algorithm and applied on top of that auxiliary logic for determining the best partitioning scheme.

II. Related Work

Mining boolean association rules over larger knowledge bases was early mentioned in [1], and later studied and presented in [3], for the case of databases with only categorical attributes.

Practically, the information in databases is not limited to categorical attributes, but also contains much quantitative data.

As mentioned before, mining quantitative association rules was introduced and an algorithm proposed in [2]. The algorithm studies the discretization of quantitative attributes domains into intervals in order to reduce the domain for a more categorical one.

The analysis of clusters is based on partitioning a set of numerical data into a number of subgroups, based on the fact that the objects within that group have certain similarities.

This approach doesn't represent all the times the real data, where boundaries between subgroups might be fuzzy and more detailed description of the objects inside clusters are required.

That is why many similar problems have been partially or totally solved in fuzzy environments. Several related papers present this approach: [4], [5], [6]. There are three major difficulties encountered during fuzzy clustering of real data:

1) the number of clusters cannot be defined apriori and optimal number has to be determined,
2) location and clusters centroids type cannot be known before and initial guess has to be made and
3) there is a great variance of cluster items that needs to be handled.

Since we are dealing with real data from a public dataset we are using, we are trying, through our work, to handle the three concerns presented above in an efficient manner.

What we are using for this is Fuzzy C-Means as clustering algorithm and related logic for determination of the optimal clusters set.

III. Numerical Attributes Clustering Using Fuzzy C-Means

A. Fuzzy Sets

Fuzzy sets theory has been initiated by an observation made in 1965 by Zadeh, saying that ”more often than not, the classes of objects encountered in the real physical world do not have precisely defined criteria of membership”.

978-83-60810-51-4/$25.00 © 2012 IEEE

The particularity of fuzzy sets is to capture the idea of partial membership. The characteristic function of a fuzzy set, often called membership function, is a function whose range is an ordered membership set containing more that two (often a continuum of) values (typically, the unit interval). Therefore, a fuzzy set is often understood as a function.

The fuzzy sets and their corresponding membership functions provided by the experts may not be suitable for decision association rules in a database. The quality of the results relies on the appropriateness of the fuzzy sets to the given data.

Some attributes have discrete nominal domain, and others have continuous numeric domain. In our study, we assume that discrete nominal domain attributes are characterized by crisp values and continuous numeric domain attributes are characterized by crisp values, interval values and fuzzy numbers.

In the case of training data, each data has the class information along with its confidence degree. In description of fuzzy values for continuous numeric attributes, trapezoidal fuzzy numbers are widely used since they can sufficiently well represent fuzzy values and they are simple to describe and process.[7]

Trapezoidal fuzzy numbers *Trap(*α*,* β*,* γ*,* δ*)* are defined as follows:

$$Trap(\alpha,\beta,\gamma,\delta) = \begin{cases} 0, if x < \alpha \\ (x-\alpha)/(\beta-\alpha), if \alpha \leq x \leq \beta \\ 1, if \beta \leq x \leq \gamma \\ (x-\delta)/(\gamma-\delta), if \gamma \leq x \leq \delta \\ 0, if x > \delta \end{cases} \tag{1}$$

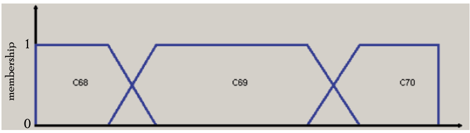

Fig. 1. Trapezoidal fuzzy numbers representation

B. Fuzzy C-Means Algorithm

Fuzzy C-Means (FCM) is a method of clustering that allows one piece of data to belong to one or more clusters. This algorithm has been developed by Dunn in 1973 and improved by Bezdek in 1981. It is based on the following objective function:

$$J_m = \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^m \|x_i - c_j\|^2, 1 \leq m \leq \infty \tag{2}$$

Fuzzy partitioning is implemented through an iterative optimization of the objective function.

The algorithm is composed of the following steps:

1) Initialize $U = [u_{ij}]$ matrix, $U^{(0)}$
2) At k-step: calculate the centers vectors $C^{(k)} = [c_j]$ with $U^{(k)}$

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^m * x_i}{\sum_{i=1}^{N} u_{ij}^m} \tag{3}$$

3) Update $U^{(k)}$, $U^{(k+1)}$

$$c_{ij} = \frac{1}{\sum_{k=1}^{C} (\frac{\|x_i - c_j\|}{\|x_i - c_k\|})^{\frac{2}{m-1}}} \tag{4}$$

4) If

$$\|U^{(k+1)} - U^{(k)}\| < \epsilon \tag{5}$$

then STOP; otherwise return to step 2.

In order to obtain good results, the initial centroids of the clusters have been generated randomly as and iteratively improved through the cycles of the algorithm. Also, as an improvement, empty clusters generated by the algorithm have been removed. This was a very useful step for further calculations.

IV. Fuzzy Sets Partitioning using Cluster Optimal Index

A very common problem in clustering is finding the optimal set of clusters that best describe the data set. Many clustering algorithms generate a required set of clusters passed as input. In order to solve this problem, the solution would be to repetitively run the algorithm with a different set of inputs until the best schema is found. In order to validate that, an auxiliary measure needs to be taken care of. We called this cluster optimal index.

A number of cluster validity indices are described in the literature. A cluster validity index for crisp (non fuzzy) clustering is proposed in [13]. An alternative has been proposed in [12]. The implementation of most of these measures is very expensive computationally, especially when the number of clusters and the number of objects in the data set grow very large. For a given attribute X, the following measures have been taken into consideration:

Variance of attribute X

$$\sigma^2(X) = \frac{1}{n} \sum_{k=1}^{N} (x_k - x)^2 \tag{6}$$

Variance of cluster i

$$\sigma^2(X_i, r_i) = \frac{\sum_{k=1}^{n_i} (x_{i_k} - r_i)^2}{n_i} \tag{7}$$

The average separation for c clusters

$$Scat(X,R) = \frac{\frac{1}{c}\sum_{i=1}^{c} \sigma^2(X_i, r_i)}{\sigma^2(X)} \tag{8}$$

$Scat(X,R)$ indicates the average compactness of clusters. A small value for this term indicates compact clusters as the scattering within clusters increases (they become less compact) the value of Scat(X,R) also increases.

Total separation between clusters

$$Dis(R) = \frac{D_{max}}{D_{min}} \sum_{i=1}^{c} (\frac{1}{\sum_{j=1}^{c} |r_i - r_j|}) \tag{9}$$

The term “total separation” sounds like a measure we want to maximize. In this case the opposite holds: a smaller value is better. $Dis(R)$ indicates the total separation between the c clusters and generally, this term will increase within the number of clusters.

Cluster optimal index has been calculated as follows:

$$OptIndex(X, R) = \alpha * Scat(X, R) + Dis(R) \quad (10)$$

considering α a factor equal to $Dis(c_{max})$ for the maximum number of input clusters.

The generalized membership function for the clusters is given by the following formula:

$$f(r_1, x) = \begin{cases} 0, if x \leq d_{i-1}^{+} \\ \frac{d_{i-1}^{+}-x}{d_{i-1}^{+}-d_i^{-}}, if d_{i-1}^{+} < x < d_i^{-} \\ 1, if d_i^{-} \leq x \leq d_i^{+} \\ \frac{d_{i+1}^{-}-x}{d_{i+1}^{-}-d_i^{+}}, if d_i^{+} < x < d_{i+1}^{-} \\ 0, if x \geq d_{i+1}^{-} \end{cases} \quad (11)$$

The steps of finding fuzzy sets can be shortly summarized as follows:

1) Finding the best clustering scheme using optimal cluster index
2) finding fuzzy sets with c clusters centers and
3) calculating the corresponding membership functions.

In details, the algorithm can be described as:

```
Main (FCM, X, $min_c$, $max_c$, p)
• Phase I: calculate the optimal number of clusters and its centroids
    – Initialize: c = $max_c$
    – Repeat
        ∗ Run FCM (clustering algorithm) for data set X to produce c cluster centers
        ∗ Calculate optimal cluster index OptIndex(X,R)
        ∗ if(c=$max_c$ then
          $\alpha = Dis(max_c)$
          $BestOptIndex = OptIndex(c)$
          $best_c = c$
        ∗ endif
        ∗ else if $(OptIndex(c) < BestOptIndex)$ then
          $best_c = c$
          $BestOptIndex = OptIndex(c)$
        ∗ endif
        ∗ c = c-1
    – until $c = min_c - 1$
• Phase II: calculate fuzzy sets with the c cluster centers
    – for i = 1 to $best_c$ do
        ∗ if $i < best_c$ then calculate $d_i^{+}$ using overlap percentage p
        ∗ if $i \geq 2$ then calculate $d_i^{-}$ using overlap percentage p
    – endfor
• Phase III: calculate membership function for each fuzzy set
    – for each $x \in X$ do
        ∗ foreach $r_i \in R$ do
        ∗ calculate membership function $f$
        ∗ endfor
    – endfor
```

V. Implementation method

The development of the application that represents the basis for our framework was done in C#.NET using .NET Framework 4.0 and the major advantages that it offers.

The decision was influenced by some of the many reasons people are using MS based technologies for their work, especially .NET Framework:

- Consistent Programming Model
- Direct Support for Security
- Simplified Development Efforts
- Easy Application Deployment and Maintenance

We've developed several auxiliary tools that helped us in evaluation and measurements.

VI. Experimental Results

A. Testing Data Set

In order to test the mechanism we build, we decided to use a public dataset providing real world data. The one we selected is called *ImageCLEF - Image Retrieval in CLEF*[8]. It contains segmented and annotated images for evaluation of automatic image annotation and for studying their impact on multimedia information retrieval.

All the images have been manually processed, segmented and annotated based on a predefined vocabulary of labels. Visual features have also been extracted from each region. We have chosen this data source because it consists a well reference for our further development. Basically, we would have consistent ground truth images for further analysis and evaluation.

The collection contains the following sets of data:

- Segmentation masks: one per region: 99.535 files; one per image: 20.000 files.
- Annotations: one per region: 99.535 regions were manually annotated.
- Spatial relationships: one per image: 20.000 files.
- Visual features: a vector of features per region: 99.535 vectors of attributes.

In 2 is displayed an example of segmented and annotated image:

B. Performance Results

We have measured the performance of the mechanism we built on the dataset that we considered for testing. Initially we calculated how is the execution time increasing if the number of attribute values is increasing but for the same number of clusters. We obtained the graphic in 3

Another set of tests have been performed on overall partitioning mechanism. We increased the elements count from 10 to 20000 and observed that for a bigger number of values, the execution time is not increasing rapidly, which makes us

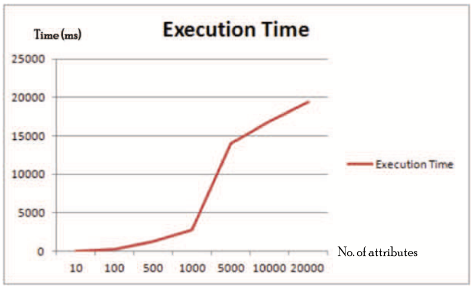

Fig. 4. Partitioning different value sets

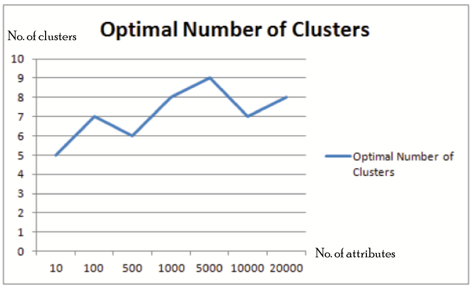

Fig. 5. Optimal number of clusters

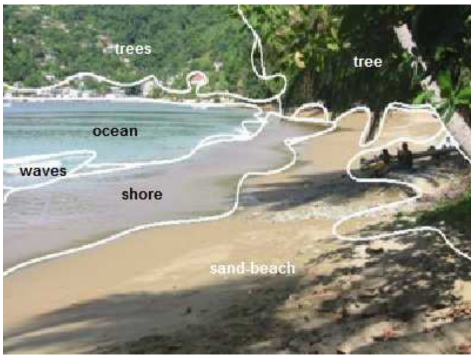

Fig. 2. Segmented and annotated image from testing dataset

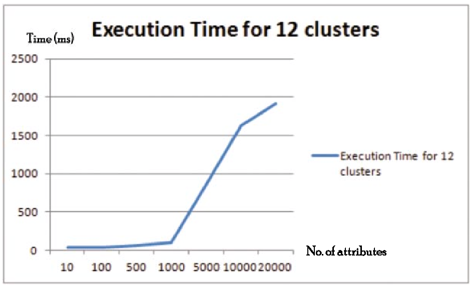

Fig. 3. Clustering execution time for 12 clusters

believe that we would have good performance on processing a large number o images. 4

The optimal number of partitions had a constant behavior, mainly depending on the values we processed. since their variance was not so big, the number of cluster didn't vary so much, which was expected.5

VII. Conclusion

As a first step from a more complex project we are developing, the resulting metrics were very satisfying for us and make us believe that we would have further good results. The dataset is very comprehensive and offers a lot of useful data for our testing.

This algorithm is quite powerful since the the merging cost evaluations requires simple identifications of complex models which is easy to implement and computationally cheap to calculate.

We propose a mechanism to find the optimal partitions as fuzzy sets based on clustering techniques. From experiments we found that the method produces meaningful results and has reasonable efficiency.

References

[1] R. Agrawal, T. Imielinski, A. Swami, *Mining association rules between sets of items in large databases* Proceedings of ACM SIGMOD, 1993

[2] R. Srikant and R. Agrawal, *Mining quantitative association rules in large relational tables* Proceedings of ACM SIGMOD, 1996

[3] R. Agrawal and R. Srikant, *Fast algorithms for mining association rules in large databases* Proceedings of the 20th VLDB Conference, 1994

[4] J. C. Bezdek, *Feature selection for binary data: Medical diagnosis with fuzzy sets* Proceedings of the 25th National Computer Conference, 1976

[5] T. Gou and B. Dubuisson *A loose-patern approach to clustering data sets* IEEE Trans. Pattern Anal. Machine Intell., 1985

[6] A.K. Jain and J.V. Moreau *Bootstrap technique in cluster analysis* Pattern Recognition, vol. 20, 1987

[7] J. Jang, *Structure determination in fuzzy modeling: A fuzzy CART approach* Proceedings IEEE Conf on Fuzzy Systems, 1994

[8] ImageCLEF - Image Retrieval in CLEF, *Segmented and annotated IAPR TC-12 dataset* http://imageclef.org/SIAPRdata

[9] A. Gyenesei, *Determining Fuzzy Sets for Quantitative Attributes in Data Mining Problems*

[10] A. Gyenesei, *Fuzzy Partitioning of Quantitative Attribute Domains by a Cluster Goodness Index* TUCS Technical Reports, 2000

[11] K.-M. Lee,K.Mi Lee, J-H. Lee, H. Lee-Kwang *A Fuzzy Decision Tree Induction Method for Fuzzy Data* IEEE International Fuzzy Systems Conference Proceedings, 1999

[12] Z. Huang, *A Fast Clustering Algorithm to Cluster very Large Categorical Data Sets in Data Mining* DMDK, 1997

[13] J. C. Dunn, *Well separated clusters and optimal fuzzy partitions* J. Cybern. Vol. 4, 1974

[14] X. L. Xie, G. Beni *A validity measure for Fuzzy Clustering* IEEE Transactions on Patern Analysis and machine Intelligence, 1991

[15] R. N. Dave, *Validating fuzzy partitions obtained through c-shells clustering* Pattern Recognition Letters, 1996

[16] R. Unnikrishnan, C. Pantofaru, and M. Hebert, *Toward Objective Evaluation of Image Segmentation Algorithms* IEEE Transactions on pattern analysis and machine inteligence, Vol. 29, No. 6, 2007.

[17] C. Odet, B. Belaroussi, and H. Benoit-Cattin, *Scalable discrepancy measures for segmentation evaluation* Proc. of the Int. Conf. on Image Processing (ICIP02), 1, 785–788, 2002

[18] D. Martin, C. Fowlkes, D. Tal, and J. Malik, *A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics* Proc. Int. Conf. Comp. Vis., vol. 2, pp. 416-425, 2001.

[19] D. Burdescu, M. Brezovan, E. Ganea, and L. Stanescu, *A New Method for Segmentation of Images Represented in a HSV Color Space*, Lecture Notes in Computer Science, 5807, 606–617, 2009.