A Visual Quality Index for Fuzzy C-Means

Aybüke Öztürk, Stéphane Lallich, and Jérôme Darmont

Université de Lyon, Lyon 2, ERIC EA 3083
5 avenue Pierre Mendès France, F69676 Bron Cedex, France
aybuke.ozturk@univ-lyon2.fr, stephane.lallich@univ-lyon2.fr,
jerome.darmont@univ-lyon2.fr

Abstract. Cluster analysis is widely used in the areas of machine learning and data mining. Fuzzy clustering is a particular method that considers that a data point can belong to more than one cluster. Fuzzy clustering helps obtain flexible clusters, as needed in such applications as text categorization. The performance of a clustering algorithm critically depends on the number of clusters, and estimating the optimal number of clusters is a challenging task. Quality indices help estimate the optimal number of clusters. However, there is no quality index that can obtain an accurate number of clusters for different datasets. Thence, in this paper, we propose a new cluster quality index associated with a visual, graph-based solution that helps choose the optimal number of clusters in fuzzy partitions. Moreover, we validate our theoretical results through extensive comparison experiments against state-of-the-art quality indices on a variety of numerical real-world and artificial datasets.

Keywords: Fuzzy Clustering, Fuzzy C-Means, Quality Indices, Visual Index, Elbow Rule

1 Introduction

Clustering refers to the assignment of unlabeled data points into clusters (groups) so that the points belonging to the same cluster are more similar to each other than those within different clusters. There are various types of clustering strategies, including crisp and fuzzy clustering. In crisp (or hard) clustering, a data point can belong to one and only one cluster, while in fuzzy clustering [1], a data point can belong to several clusters. Fuzzy clustering is very useful in many applications, e.g., the text categorization of various news into different clusters: a science, a business, and a sport cluster; where an article containing the keyword "gold" could belong to all three clusters. Furthermore, it is also possible to open discussions with domain experts when using fuzzy clustering.

Clustering algorithms behave differently for different reasons. The first reason relates to dataset features such as geometry and the density distribution of clusters. The second reason is the choice of input parameters such as the fuzziness coefficient m ($m = 1$ indicating that clustering is crisp and $m > 1$ that clustering becomes fuzzy).

These parameters all affect the quality of clustering. To study how the choice of parameters impacts clustering quality, we need a quality criterion. For instance, when the dataset is well separated and has only two variables, a scatter plot can help determine the number of clusters. However, when the dataset has more than two variables, a good quality index is needed to compare various cluster configurations and choose the appropriate number of clusters.

Achieving a good clustering involves both minimizing intra-cluster distance (compactness) and maximizing inter-cluster distance (separability). A common issue in this process is that clusters are split up while they could be more compact. Many cluster quality indices have been proposed to address this problem for hard and fuzzy clustering, but none of them is always highly efficient [2].

Moreover, there is no real-life golden standard for clustering analysis, since various experts may have different points of views about the same data and express different constraints on the number and size of clusters. Thanks to a visual index, different solutions can be presented with respect to the data. Thus, experts can make a trade-off between their opinion and the best local solutions proposed by the visual index.

Hence, in this paper, we first review existing quality indices that are well-suited to fuzzy clustering, such as [3,4,5,6,7,8]. Then, we propose an innovative, visual quality index for the well-known Fuzzy C-Means (FCM) method. Moreover, we compare our proposal with state-of-the-art quality indices from the literature on several numerical real-world and artificial datasets.

The remainder of this paper is organized as follows. Section 2 recalls the principles of fuzzy clustering. Section 3 surveys quality indices for fuzzy clustering. Section 4 details our visual quality index. Section 5 reports on the experimental comparison of our quality index against existing ones on different datasets. Finally, we conclude this paper and provide research perspectives in Section 6.

2 Principles of Fuzzy Clustering

Fuzzy inertia is a core measure in fuzzy clustering. Fuzzy inertia FI (Equation 1) is composed of fuzzy within-inertia FW (Equation 2) and fuzzy between-inertia FB (Equation 3). Membership coefficients u_{ik} of data point i to cluster k are usually stored in a membership matrix U that is used to calculate FW, FB and FI. Note that $FI = FW + FB$. Moreover, FI is not constant because it depends on u_{ik}. When FW changes, the values of FI and FB also change.

$$FI = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^{m} d^2(x_i, \overline{x}) \tag{1}$$

$$FW = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^{m} d^2(x_i, c_k) \tag{2}$$

$$FB = \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^{m} d^2(c_k, \overline{x}) \tag{3}$$

where n is the number of instances, K is the number of clusters, m is the fuzziness coefficient (by default, $m = 2$), c_k is the center of the k^{th} cluster $\forall 1 \leq k \leq K$, $\overline{x}$ is the grand mean (the arithmetic mean of all data – Equation 4), and function $d^2()$ computes the squared Euclidean distance.

$$\overline{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{4}$$

FCM is a common method for fuzzy clustering that adapts the principle of the K-Means algorithm [9]. FCM, proposed by [10] and extended by [11], applies on numerical data. Since numerical data are the most common case, we choose to experiment our proposals with FCM.

The aim of the FCM algorithm is to minimize FW. It starts by choosing K data points as initial centroids of the clusters. Then, membership matrix values u_{ik} (Equation 5) are assigned to each data point in the dataset. Centroids of clusters c_k are updated based on Equation 6 until a termination criterion is reached successfully. In FCM, this criterion can be a fixed number of iterations t, e.g., $t = 100$. Alternatively, a threshold ϵ can be used, e.g., $\epsilon = 0.0001$. Then, the algorithm stops when $FW_{K+1}/ \mid FW_{K+1}$ - $FW_K \mid < \epsilon$.

$$u_{ik} = \frac{1}{\sum_{j=1}^{K} (\frac{\|x_i - c_k\|^2}{\|x_i - c_j\|^2})^{\frac{1}{m-1}}} \tag{5}$$

$$c_k = \frac{\sum_{i=1}^{n} u_{ik}^m x_i}{\sum_{i=1}^{n} u_{ik}^m} \tag{6}$$

3 Fuzzy Clustering Quality Indices

According to Wang et al. [12], there are two groups of quality indices. Quality indices in the first group are based only on membership values. They notably include partition coefficient index V_{PC} [3] (Equation 7; $\frac{1}{K} \leq V_{PC} \leq 1$; to be maximized) and Chen and Linkens' index V_{CL} [4] (Equation 8; $0 \leq V_{CL} \leq 1$; to be maximized). V_{CL} takes into consideration both compactness (first term of V_{CL}) and separability (second term of V_{CL}).

$$V_{PC} = \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} u_{ik}^2 \tag{7}$$

$$V_{CL} = \frac{1}{n} \sum_{i=1}^{n} max_k(u_{ik}) - \frac{1}{c} \sum_{k=1}^{K-1} \sum_{j=k+1}^{K} \left[\frac{1}{n} \sum_{i=1}^{n} min(u_{ik}, u_{ij}) \right], \tag{8}$$

where $c = \sum_{k=1}^{K-1} k$.

Quality indices in the second group associate membership values to cluster centers and data. They include an adaptation of the Ratio index V_{FRatio} to fuzzy

clustering [5] (Equation 9; $0 \leq V_{FRatio} \leq +\infty$; to be maximized), Fukuyama and Sugeno's index V_{FS} [6] (Equation 10; $-FI \leq V_{FS} \leq FI$; to be minimized), and Xie and Beni's index V_{XB} [7,13] (Equation 11; $0 \leq V_{XB} \leq FI/n{*}min\|x_j - v_k\|^2$; to be minimized).

$$V_{FRatio} = FB/FW \tag{9}$$

$$V_{FS} = FW - FB \tag{10}$$

$$V_{XB} = \frac{\sum_{k=1}^{K}\sum_{i=1}^{n} u_{ik}^{m}\|x_i - v_k\|^2}{n * min_{j,k}\|v_j - v_k\|^2} \tag{11}$$

When the number of clusters increases, the value of quality indices mechanically increases, too. Then, the important question is: how useful is the addition of a new cluster? To answer this question, the most common solutions are penalization and the Elbow Rule [14].

The first way to penalize a quality index is to multiply it by a quantity that diminishes the index when the number of clusters increases. In this case, the main difficulty is to choose the penalty. For instance, the penalized version of V_{FRatio} is Calinski's V_{FCH} [5] (Equation 12; $0 \leq V_{FCH} \leq +\infty$; to be maximized), where the penalty is based on both the number of clusters and data points.

$$V_{FCH} = \frac{FB/(K-1)}{FW/(n-K)} = \frac{n-K}{K-1}\frac{FB}{FW} \tag{12}$$

The second way to penalize a quality index is to evaluate index evolution relatively to the number of clusters, by considering the curve of the index' successive values. The most appropriate value of K can be determined visually by help of the Elbow Rule or algebraic calculation [15].

To construct a visual determination of the Elbow Rule, K is represented on the horizontal axis and the considered quality index on the vertical axis. Then, we look for the value of K where there is a change in the curve's concavity. This change represents the optimal number of clusters K. To construct an algebraic determination, let i_K being the index value for K clusters. The variation of i_K before K and after K are compared. In case of a positive Elbow, the second difference $min_K((i_{K+1} - i_K) - (i_K - i_{K-1}))$ is minimized. Yet, since the values before K and after K are used for calculation, the Elbow Rule can be applied to more than two clusters only.

Among all the above-stated quality indices, there is no single quality index that gives the best result for any dataset. Thus, there is room for a new quality index that is specifically tailored for fuzzy validation and helps the user choose the value of K.

4 An Index Associated with a Visual Solution

Building a new quality index, we first consider FW to evaluate compactness and FB to evaluate separability. We can choose to calculate either $FB - FW$, which

is similar to V_{FS} except for the sign, or $FB \div FW$, which is similar to V_{FRatio}. Unfortunately, $FI = FB+FW$ is not constant and $FB-FW \in [-FI, +FI]$. To take this particularity of fuzzy clustering into account, we propose to standardize $FB - FW$ by considering the *Standardized Fuzzy Difference* $SFD = (FB - FW) \div FI$ instead. Then, $SFD \in [-1, +1]$.

Adding a new cluster often improves clustering quality mechanically. Thus, many authors penalize the quality index with respect to K (the smaller n is, the greater the penalty), e.g., V_{FCH} (Section 3). To obtain a penalized index, SFD is first linearly transformed in an index belonging to $[0, 1]$, obtaining the *Transformed Standardized Fuzzy Difference* $TSFD$ (Equation 13; $TSFD \in [0, 1]$; to be maximized). Finally, by penalizing $TSFD$ as V_{FCH}, we obtain the *Penalized Standardized Fuzzy Difference* $PSFD$ (Equation 14; $PSFD \in [0, (n - K)/(K - 1)]$; to be maximized).

$$TSFD = \frac{1 + SFD}{2} = \frac{FB}{FI} \tag{13}$$

$$PSFD = TSFD * \frac{n - K}{K - 1} = \frac{FB - FW}{FI} * \frac{n - K}{K - 1} \tag{14}$$

Instead of penalizing the quality index, another solution is to visualize the search for the best number of clusters K. First solution is to apply the Elbow Rule to $TSFD$. $TSFD$ is plotted with respect to K in Figure 1(a). The drawback of this method is that the horizontal axis corresponds to an arithmetic scale of K values, which is not satisfying. To fix this problem, we suggest to plot FB with respect to FI, which we call *Visual* $TSFD$. Our aim is not to give an automatic solution, but to help the user visually choose the most appropriate K value. The visualization we propose is shown in Figure 1(b), where the blue line plots $TSFD$ with respect to K, the full red line is the diagonal that corresponds to the best solutions ($FB = FI$) such that $TSFD = 1$, and the dashed red line connects the origin to each point associated with K values. The smaller the angle between the full red line and the dashed red line, the better is the solution. As the value of K increases, the angle between the dashed red line and the diagonal decreases. Then, we choose the value of K beyond which the decrease becomes negligible. This value is considered as the optimal number of clusters. For example, in Figure 1(b), a first solution could be $K = 4$, a better solution $K = 6$, and it is not very interesting to consider $K > 6$.

5 Experimental Validation

In this section, we compare our proposals $TSFD$, $PSFD$, *Visual TSFD* and the use of the Elbow Rule to state-of-the-art clustering quality indices for FCM-like clustering algorithms, i.e., V_{PC}, V_{CL}, V_{FCH}, V_{FS} and V_{XB} (Section 3).

In our experiments, the FCM algorithm is parameterized with its default settings: termination criterion $\epsilon = 0.0001$ and default fuzziness coefficient $m = 2$. All clustering quality indices are coded in Python version 2.7.4.

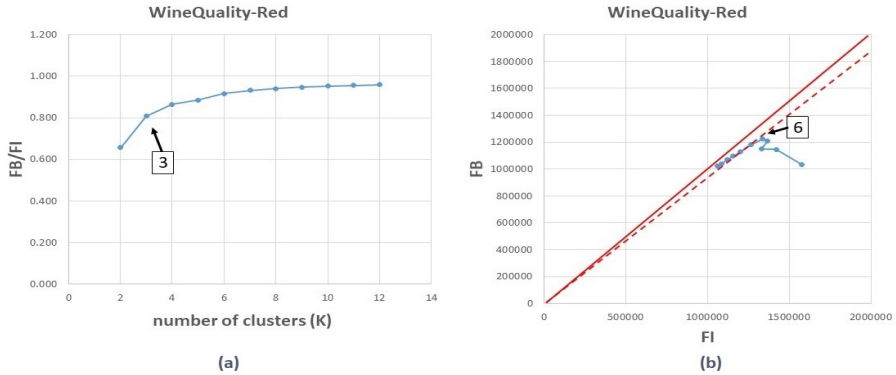

Fig. 1: Comparison of Elbow Rule (a) and *Visual TSFD* (b) on the WineQuality-Red dataset (Table 1)

5.1 Datasets

Quality indices are compared on ten real-life datasets (Table 1; IDs 1-10) from the UCI Machine Learning Repository[1] and seven artificial datasets (Table 1; IDs 11-17). In real-life datasets, the true number of clusters is assimilated to the number of labels. Although using the number of labels as the number of clusters is debatable, it is acceptable if the set of descriptive variables explains the labels well. In artificial datasets, the number of clusters is known by construction. Moreover, we created new artificial datasets by introducing overlapping and noise to some of the existing datasets, such as E1071-3 [16], Ruspini [1] and E1071-5 [16] (Table 1; IDs 12-14). To create a new dataset, new data points are introduced, and each must be labeled. To obtain a dataset with overlapping, we modify the construction of the E1071 artificial datasets [16]. In the original datasets, there are three or five clusters of equal size (50). Cluster i is generated according to a Gaussian distribution $N(i; 0.3)$. To increase overlapping in the three clusters while retaining the same cluster size, we only change the standard deviation from 0.3 to 0.4. Then, there is no labeling problem. To introduce noise in a dataset, we add in each cluster noisy points generated by a Gaussian variable around each label gravity center. Noisy data are often generated by distributions with positive skewness. For example, in a two-dimensional dataset, for each label, we add points that are far away from the corresponding gravity center, especially on the right-hand side, which generally contains the most points. Then, we draw a random number r between 0 and 1. If $r \leq 0.25$, the point is attributed to the left-hand side. Otherwise, the point is attributed to the right-hand side. This method helps obtain noisy data that are $^1/_4$ times smaller and $^3/_4$ times greater, respectively, than the expected value for the considered label. This process is applied to the Ruspini dataset [1].

[1] http://archive.ics.uci.edu/ml/

5.2 Experimental Results

In our experiments, all validation indices (Sections 3 and 4) are applied on all the datasets from Table 1. Moreover, since presenting all the results would take too much space, we retain only the best results for each index (even excluding $PSFD$).

Table 1: Quality Indices Experiment Results with Different Datasets

ID	Datasets	# of data points	# of clusters	V_{PC}	V_{CL}	FB	V_{FCH}	V_{FS}	V_{XB}	Elbow V_{TSFD}	Visual V_{TSFD}
1	Wine	178	3	2	2	8	12	8	2	**3**	5
2	Iris	150	3	2	2	**3**	**3**	**3**	2	**3**	**3**
3	Seeds	210	3	2	**3**	**3**	**3**	**3**	2	**3**	**3**
4	Glass	214	6	2	2	12	12	12	2	4	5,7
5	Vehicle	846	4	2	2	2	2	5	2	3	**4**,5
6	Segmentation	2310	7	2	4	4	4	12	12	3	**7**,8
7	Movement Libras	360	15	2	18	16	16	18	2	14	14,16
8	Ecoli	336	8	2	3	3	3	12	3	3	3,7
9	Yeast	1484	10	2	2	5	2	12	2	4	7,8
10	WineQuality-Red	1599	6	2	2	**6**	7	**6**	2	3	**6**
11	Bensaid [17]	49	3	**3**	**3**	9	11	11	**3**	**3**	5
12	E1071-3 [16]	150	3	**3**	**3**	**3**	**3**	**3**	**3**	**3**	**3**
13	Ruspini [1]	75	4	**4**	**4**	**4**	**4**	**4**	**4**	3	**4**
14	E1071-5 [16]	250	5	2	**5**	4	**5**	**5**	2	3	**5**
15	E1071-3-overlapped	150	3	2	**3**	**3**	2	**3**	2	**3**	**3**
16	Ruspini_noised	95	4	**4**	12	**4**	**4**	**4**	**4**	**4**	**4**
17	E1071-5-overlapped	250	5	2	2	4	**5**	4	2	3	**5**
# of wins for real-life datasets				0	1	3	2	3	0	3	**5**
# of wins for artificial datasets				4	5	4	5	5	4	4	**6**
Total # of wins				4	6	7	7	8	4	7	**11**

As shown in Table 1, it is more difficult to predict an appropriate number of clusters for real-life datasets than for artificial datasets. Considering all indices, the average rate of success is indeed 21% in the case of real data, against 66% in the case of artificial data. Whatever the type of data, *Visual TSFD* outperforms the other indices, with 5 wins out of 10 in the case of real datasets, and 6 wins out of 7 in the case of artificial datasets. The worst results are obtained with V_{PC} and V_{XB} (0/10 and 4/7 wins each). The other indices achieve intermediary results. In addition, when the value given by *Visual TSFD* is erroneous, it is quite close to the expected K, in contrast to V_{FS}, our closest competitor (Table 1; Wine, Glass, Segmentation, Ecoli and Bensaid). For example, the optimal number of clusters should be 6 for the Glass dataset. $V_{FS} = 12$, *Visual TSFD*'s results

are 5 and 7. Furthermore, we compare in Figures 2 and 3 *Visual TSFD* and the plot obtained with the Elbow Rule (which is labeled Elbow $TSFD$) with respect to K, on a sample of both real-life and artificial datasets bearing different characteristics, i.e., Glass, Vehicle, Ecoli, Ruspini, Ruspini_noised and E1071-5-overlapped (Table 1). As is clearly visible from Figures 2 and 3, *Visual TSFD* gives a better visual idea than Elbow $TSFD$. Elbow $TSFD$ indeed highlights K values of 3 or 4, while the $TFSD$ blue plot systematically indicates larger K values.

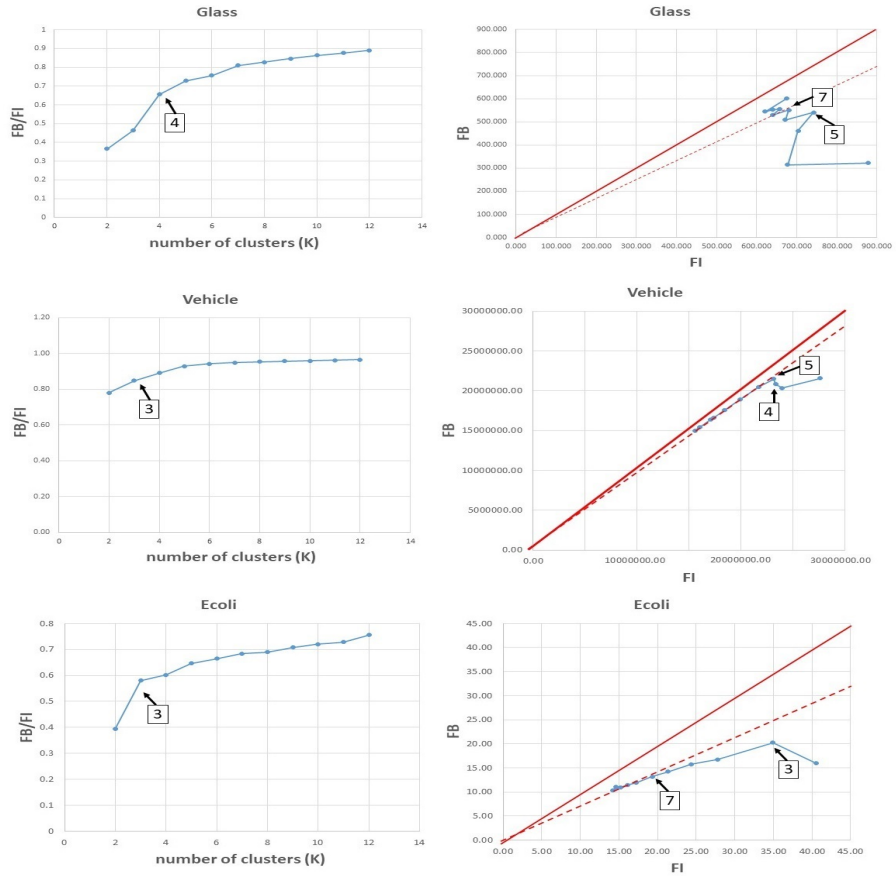

Fig. 2: Comparison of Elbow $TSFD$ and *Visual TSFD* (1/2)

Eventually, since our work aims at real-life datasets, there is no ground truth or golden standard for clustering analysis. In such a context, *Visual TSFD* has the advantage of providing options to experts instead of outputting a single K

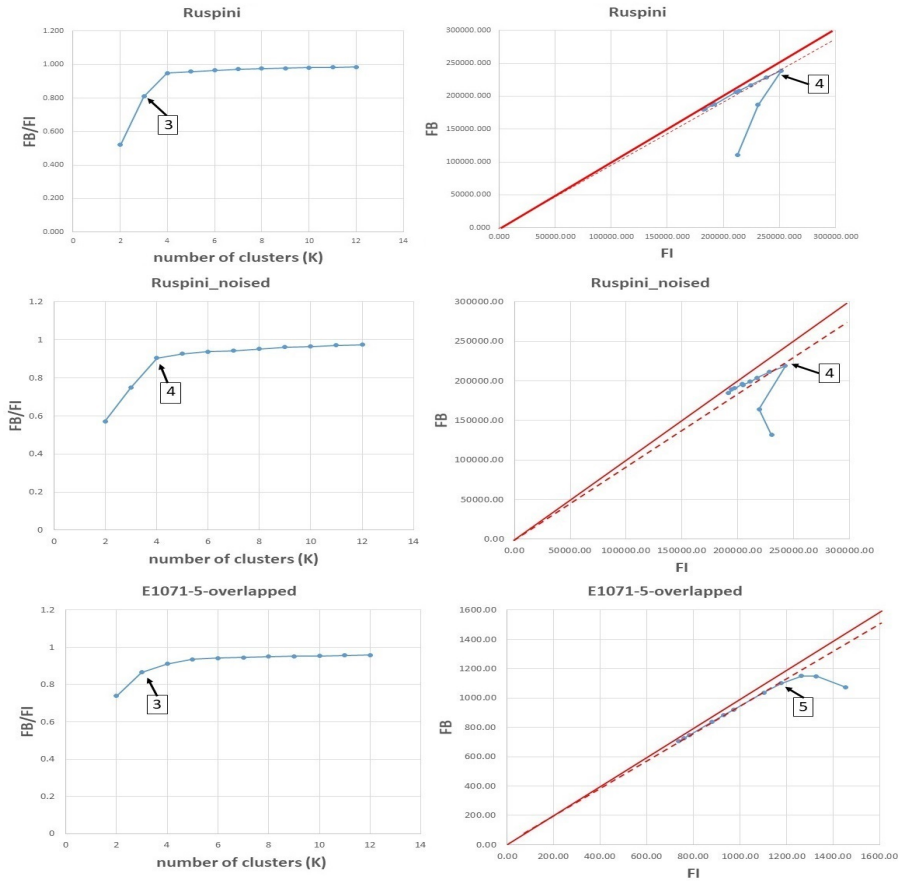

Fig. 3: Comparison of Elbow $TSFD$ and *Visual TSFD* (2/2)

value. This makes our method more flexible that the existing ones in real-life scenarios.

6 Conclusion and Perspectives

In this paper, we propose a novel quality index for FCM called *Visual TSFD*, which provides an overview of fuzzy clustering with respect to the number of clusters. We compare *Visual TSFD* to several clustering quality methods from the literature and experimentally show that it outperforms existing methods on various datasets. Furthermore, *Visual TSFD* can also be used in the case of categorical data with Fuzzy K-Medoids [18]. Thus, *Visual TSFD* allows to deal with heterogeneous datasets, which makes our method a simple but noteworthy contribution, in our opinion. As a result, our next step is to design an ensem-

ble fuzzy clustering method based on *Visual TSFD* that would deal with both numerical and categorical data.

Acknowledgments

This project is supported by the Rhône Alpes Region's ARC 5: "Cultures, Sciences, Sociétés et Médiations" through A. Öztürk's Ph.D. grant.

References

1. Ruspini, E.H.: Numerical methods for fuzzy clustering. Information Sciences **2**(3) (1970) 319–350
2. Pal, N.R., Bezdek, J.C.: Correction to" on cluster validity for the fuzzy c-means model"[correspondence]. IEEE transactions on fuzzy systems **5**(1) (1997) 152–153
3. Bezdek, J.C.: Cluster validity with fuzzy sets. (1973)
4. Chen, M.Y., Linkens, D.A.: Rule-base self-generation and simplification for data-driven fuzzy models. In: Fuzzy Systems, 2001. The 10th IEEE International Conference on. Volume 1., IEEE (2001) 424–427
5. Caliński, T., Harabasz, J.: A dendrite method for cluster analysis. Communications in Statistics-theory and Methods **3**(1) (1974) 1–27
6. Fukuyama, Y., Sugeno, M.: A new method of choosing the number of clusters for the fuzzy c-mean method. In: Proc. 5th Fuzzy Syst. Symp., 1989. (1989) 247–250
7. Xie, X.L., Beni, G.: A validity measure for fuzzy clustering. IEEE Transactions on pattern analysis and machine intelligence **13**(8) (1991) 841–847
8. Zhang, D., Ji, M., Yang, J., Zhang, Y., Xie, F.: A novel cluster validity index for fuzzy clustering based on bipartite modularity. Fuzzy Sets and Systems **253** (2014) 122–137
9. MacQueen, J., et al.: Some methods for classification and analysis of multivariate observations. In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Volume 1., Oakland, CA, USA. (1967) 281–297
10. Dunn, J.C.: A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. (1973)
11. Bezdek, J.C., Ehrlich, R., Full, W.: Fcm: The fuzzy c-means clustering algorithm. Computers & Geosciences **10**(2-3) (1984) 191–203
12. Wang, W., Zhang, Y.: On fuzzy cluster validity indices. Fuzzy sets and systems **158**(19) (2007) 2095–2117
13. Pal, N.R., Bezdek, J.C.: On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy systems **3**(3) (1995) 370–379
14. Cattell, R.B.: The scree test for the number of factors. Multivariate behavioral research **1**(2) (1966) 245–276
15. Dimitriadou, E., Dolničar, S., Weingessel, A.: An examination of indexes for determining the number of clusters in binary data sets. Psychometrika **67**(1) (2002) 137–159
16. Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., Leisch, F., Chang, C.C., Lin, C.C., Meyer, M.D.: Package e1071. Version 1.6-8 (2017)
17. Bensaid, A.M., Hall, L.O., Bezdek, J.C., Clarke, L.P., Silbiger, M.L., Arrington, J.A., Murtagh, R.F.: Validity-guided (re) clustering with applications to image segmentation. IEEE Transactions on Fuzzy Systems **4**(2) (1996) 112–123
18. Park, H.S., Jun, C.H.: A simple and fast algorithm for k-medoids clustering. Expert systems with applications **36**(2) (2009) 3336–3341