Clustering Quality Evaluation Based on Fuzzy FCA

Minyar Sassi[1], Amel Grissa Touzi[1], and Habib Ounelli[2]

[1] Ecole Nationale d'Ingénieurs de Tunis
Bp. 37, Le Belvédère 1002 Tunis, Tunisia
{minyar.sassi,amel.touzi}@enit.rnu.tn
[2] Faculté des Sciences de Tunis
Campus Universitaire -1060 Tunis, Tunisia
habib.ounelli@fst.rnu.tn

Abstract. Because clustering is an unsupervised procedure, clustering results need be judged by external criteria called validity indices. These indices play an important role in determining the number of clusters in a given dataset. A general approach for determining this number is to select the optimal value of a certain cluster validity index. Most existing indices give good results for data sets with well separated clusters, but usually fail for complex data sets, for example, data sets with overlapping clusters. In this paper, we propose a new approach for clustering quality evaluation while combining fuzzy logic with Formal Concept Analysis based on concept lattice. We define a formal quality index including the separation degree and the overlapping rate.

Keywords: Clustering Quality, Overlapping Rate, Separation Degree, Validity Index, Formal Concept Analyis, Fuzzy Concept Lattice.

1 Introduction

Fuzzy clustering allows objects of a data set to belong to several clusters simultaneously, with different degrees of membership. The data set is thus partitioned into a number of fuzzy partitions (clusters) [1].

Despite being a very effective technique, difficulties arise when evaluating the quality of clusters.

So, evaluating the quality of the clustering results is an important issue in cluster analysis. Because clustering is an unsupervised procedure, clustering results need be judged by an external criterion.

For low dimensional data sets (1-, 2- or 3-dimensional), humans can also evaluate the clustering results by visual observation. For high dimensional data sets (more then 3-dimentional), there is no objective criterion for evaluating the clustering results; they are assessed using a cluster validity index.

Depending on the type of clustering approach (crisp or fuzzy), there are various validity indices designed for evaluating the clustering results [2]. The general principle of these indices consists on minimizing the compactness within a cluster and maximizing the separation between clusters.

R. Wagner, N. Revell, and G. Pernul (Eds.): DEXA 2007, LNCS 4653, pp. 639–649, 2007.

These measures play an important role in determining the number of clusters. It is expected that the optimal value of the cluster validity index should be obtained at the true number of clusters. A general approach for determining the number of clusters is to select the optimal value of a certain cluster validity index. Whether a cluster validity index yields the true number of clusters is a criterion for the validity index.

Most existing indices give good results for data sets with well separated clusters, but usually fail for complex data sets, for example, data sets with overlapping clusters. One of the main reasons for this problem is that many fuzzy clustering methods fail to distinguish between partially overlapped clusters [3].

Because they disregard lack of considering the theoretical characterization of the overlapping phenomenon, they often yield questionable results for cases involving overlapping clusters [4].

To cure this problem, we propose to use conceptual scaling theory [5] based on an extension of Formal Concept Analysis (FCA) [6] which permits us to:

- Visualizing the clusters results will help us in interpreting and distinguishing overlapping clusters, and hence,
- Evaluating the quality of clusters while calculating a separation degree and an overlapping rate for a given clustering.

The rest of the paper is organized as follows. Section 2 discusses the backgrounds in FCA based on concepts lattices and Conceptual scaling. Section 3 presents our quality evaluation process. Section 4 concludes the paper and gives some future works.

2 Backgrounds

FCA provides a conceptual framework for structuring, analyzing and visualizing data, in order to make them more understandable [6]. In FCA, application domains are organized and structured according to concept lattices. In this section, we discuss about concept lattices and conceptual scaling.

2.1 Concept Lattices

The reason for the introduction of FCA was to relate the mathematically oriented theory of lattices and orders to practical problems [6,7].

In 1979, Wille [6] recognized that this description could be formalized by the introduction of 'formal concepts' of a given data table, which consists of a set G of object, a set M of attributes and a binary relation $I \subseteq G \times M$. Then the triple $K = (G, M, I)$ is called a formal context, representing just a set of statements of the form 'object g has attribute m', written '$g\ I\ m$'.

The basic definition of a 'formal concept' of K is based on two well-known operations: For any subset $X \subseteq G$ we are interested in the set $X\uparrow$ of all common attributes of X, defined formally by $X\uparrow := \{m \in M \mid \forall g \in X \quad g\ I\ m\}$ and dually for any $Y \subseteq M$ we are interested in the set $Y\downarrow$ of all common objects of Y, defined formally by $Y\downarrow := \{g \in G \mid \forall m \in Y \quad g\ I\ m\}$. A formal concept of a formal context K is a pair (A,B) where $A \subseteq G$, $B \subseteq M$ and $A\uparrow = B$ and $B\downarrow = A$. A is called the extent, B the intent of (A,B).

The set of all formal concepts of K is denoted by $B(K)$. The conceptual hierarchy among concepts is defined by set inclusion: For (A_1,B_1), $(A_2,B_2)\in B(K)$ let $(A_1,B_1)\le(A_2,B_2)\Leftrightarrow A_1\subseteq A_2$ (which is equivalent to $B_2\subseteq B_1$.

An important role is played by the object concepts $\gamma(g):=(\{g\}\uparrow\downarrow,\{g\}\uparrow)$ for $g\in G$ and dually the attribute concepts $\mu(m):=(\{m\}\downarrow,\{m\}\downarrow\uparrow)$ for $m\in M$.

The pair $(B(K),\le)$ is an ordered set, i.e., $\le$ is reflexive, anti-symmetric, and transitive on $B(K)$.

2.2 Conceptual Scaling

An arbitrary ternary relation on a set G of 'objects' is a special case of a ternary relation among three sets of objects. In formal descriptions of measurements by data tables the following three sets play a fundamental role: A set G of 'objects', a set M of 'measurements' and a set W of values which are related by a ternary relation whose elements $(g,m,w)$are interpreted as 'object g has at measurement m the value w'. That leads to the following definition of a many-valued context (G,M,W,I) as a quadruple of four sets, where the elements of G are called 'objects', the elements of M 'many-valued attributes', the elements of W 'values', and I is a ternary relation, $I\subseteq G\times M\times W$, such that for any $g\in G$, $m\in M$ there is at most one value w satisfying $(g,m,w)\in I$. Therefore, a many-valued attribute m can be understood as a (partial) function, and we write $m(g)=w$ *iff* $(g,m,w)\in I$. A many-valued attribute m is called complete *iff* for any $g\in G$ there is (exactly one) $w\in W$ such that $m(g)=w$. (G,M,W,I) is called complete if each $m\in M$ is complete [7].

The central process in conceptual scaling theory is the construction of a formal context $S_m=(W_m,M_m,I_m)$ for each $m\in M$ such that $W_m\ni_m G:=\{m(g)\mid g\in G\}$. Such formal contexts, called conceptual scales, represent a contextual language about the set of values of m. Usually one chooses W_m as the set of all 'possible' values of m with respect to some purpose. Each attribute $n\in M_m$ is called a scale attribute. The set $n\downarrow=\{w\mid wI_m n\}$ is the extent of the attribute concept of n in the S_m scale. Hence, the choice of a scale induces a selection of subsets of W_m. The set of all intersections of these subsets constitutes just the closure system of all extents of the concept lattice of S_m.

The granularity of the language about the possible values of m induces in a natural way a granularity on the set G of objects of the given many-valued context, since each object g is mapped via m onto its value $m(g)$ and $m(g)$ is mapped via the object concept mapping γ_m of S_m onto $\gamma_m(m(g)):g\to m(g)\to\gamma_m(m(g))$.

Hence the set of all object concepts of S_m plays the role of a frame within which each object of G can be embedded. For two attributes m, $m'\in M$ each object g is mapped onto the corresponding pair: $g\to(m(g),m'(g))\to(\gamma_m(m(g)),\gamma_{m'}(m'(g)))\in B(S_m)\times B(S_{m'})$.

The standard scaling procedure, called plain scaling, constructs from a scaled many-valued context $((G,M,W,I),(S_m \mid m \in M))$, consisting of a many-valued context (G,M,W,I) and a scale family $(S_m \mid m \in M)$ the derived context, denoted by $K := (G, \{(m,n) \mid m \in M, n \in M_m\}, J)$, where $gJ(m,n)$ *iff* $m(g) I_m n$ $(g \in G, m \in M, n \in M_m)$.

The concept lattice $B(K)$ can be (supremum-) embedded into the direct product of the concept lattices of the scales [8]. That leads to a very useful visualization of multidimensional data in so-called nested line diagrams [9]).

3 The Quality Evaluation Process

As we have mention in section 1, evaluating the quality of clusters is an important issue in cluster analysis. It often based on a clustering validity index. The general principle of these indices consists on minimizing the compactness within a cluster and maximizing the separation between clusters. Most existing criteria give good results for data sets with well separated clusters, but usually fail for complex data sets, for example, data sets with overlapping clusters.

In this paper, we use conceptual representation of clustering results which permits us to formally calculate the compactness and the separation degrees which permits us to evaluate the quality of clusters. However, there are many situations in which uncertainty information also occurs. For example, it is sometimes difficult to judge whether an object belongs totally to an attribute or not. Traditional conceptual representation is hardly able to represent such vague information. To tackle this problem, we propose to combine fuzzy logic [10] with FCA as Fuzzy FCA (FFCA). Once this structure is built, we calculate a certain similarity distance based on membership degrees. This distance permits us to evaluate the compactness and the separation of the clustering result.

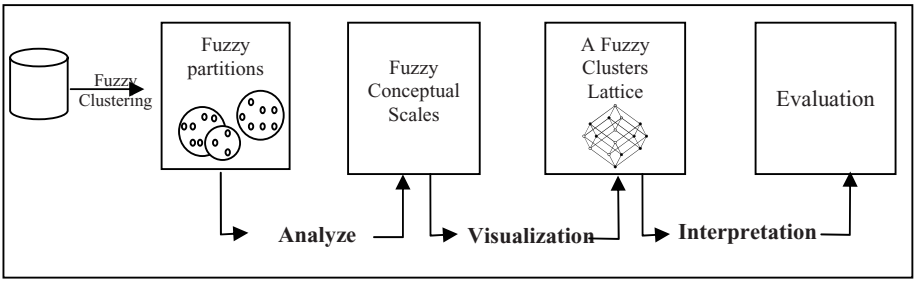

Fig. 1. The Quality Evaluation Process

The principle of our quality evaluation process determines three steps. The first step consists of analysing the fuzzy clusters for a given dataset based on fuzzy conceptual scaling. The second step consists of visualizing the results based on fuzzy

Formal Concept Analysis. This allows deducing overlapping between clusters. The third step consists of evaluating the quality of clustering results which includes the separation between clusters and compactness within a cluster. Fig. 1 shows the proposed approach.

3.1 Analyze

Fuzzy clustering methods allow objects to belong to several clusters simultaneously, with different degrees of membership. A data, set X is thus partitioned into C fuzzy partitions (clusters). In many applications training data relates individual objects to attributes that take on several values. For the generation of fuzzy formal context, we propose to relate objects with the clusters of each attribute that take on several values. These values represent the membership degrees of each object in each cluster. Fuzzy formal context incorporate fuzzy clustering, to represent vague information.

Definition 1. A fuzzy conceptual scale for a set $Y \subseteq M$ is a (single-valued) fuzzy formal context $S_Y := (G_Y, M_Y, I_Y = \varphi(G_Y \times M_Y))$ with $G_Y \subseteq \times_{m \in Y} W_m$.
The idea is to allow objects G to belong to several clusters simultaneously. We replace the attribute values in W_m with different degrees of membership. Each relation $(g,m) \in I_Y$ has a membership value $\mu(g,m)$ in $[0,1]$. The sum of the values of each fuzzy conceptual scale is equal to 1.

Definition 2. Given a fuzzy conceptual scale $S_Y := (G_Y, M_Y, I_Y = \varphi(G_Y \times M_Y))$, we define $\alpha - Cut(S_i) = (C(S_i))^{-1}$ where $C(S_i)$ is the number of clusters of scale S_i .

Example: Table. 1 present the results of fuzzy clustering applied to price and surface scales. For price scale, fuzzy clustering generate three clusters (C1,C2 and C3) for surface attribute, two clusters (C4,C5). Table 1 shows the fuzzy conceptual scales for price and surface attributes with $\alpha - Cut$. In this example, $\alpha - Cut\,(price) = 0.3$ and $\alpha - Cut\,(surafce) = 0.5$.

Table 1. Fuzzy Conceptual Scales with $\alpha - Cut$ for price and surface attributes

	Price			Surface	
	C1	C2	C3	C4	C5
A1	-	0.5	0.4	0.5	0.5
A2	0.3	0.6	-	-	0.6
A3	0.7	-	-	0.7	-
A4	-	0.4	0.5	-	0.8
A5	-	0.4	0.4	0.6	-
A6	0.5	0.3	-	0.5	0.5

3.2 Visualization

Traditional FCA is hardly able to represent fuzzy properties from uncertainly data. To tackle this problem, we use a new technique that incorporates fuzzy logic into FCA as Fuzzy Formal Concept Analysis (FFCA), in which uncertainty information is directly represented by a real number of membership value in the range of [0,1]. So we give some defined the so called Fuzzy Formal Context, the Fuzzy Formal Concept Analysis and the similarity concept.

Definition 3. Given a fuzzy formal context $K=(G,M,I)$ and an $\alpha - Cut$, we define $X^*=\{m\in M \mid \forall g\in X: \mu(g,m)\geq \alpha - Cut\}$ for $X\subseteq G$ and $Y^*=\{g\in G \mid \forall m\in Y: \mu(g,m)\geq \alpha - Cut\}$ for $Y\subseteq M$. A fuzzy formal concept (or fuzzy concept) of a fuzzy formal context (G,M,I) with an $\alpha - Cut$ is a pair $(X_f = \varphi(X), Y)$ where $X\subseteq G$, $Y\subseteq M$, $X^*=Y$ and $Y^*=X$. Each object $g\in\varphi(X)$ has a membership μ_g defined as $\mu_g=\min_{m\in Y}\mu(g,m)$. Where $\mu(g,m)$ is the membership value between object g and attribute m, which is defined in I. Note that if $Y=\{\}$ then $\mu_g=1$ for every g.

Generally, we can consider the attributes of a formal concept as the description of the concept. Thus, the relationships between the object and the concept should be the intersection of the relationships between the objects and the attributes of the concept. Since each relationship between the object and an attribute is represented as a set of membership values in fuzzy formal context, then the intersection of these membership values should be the minimum of these membership values, according to fuzzy theory [8].

Definition 4. Let (A_1,B_1) and (A_2,B_2) be two fuzzy concepts of a fuzzy formal context $K=(G,M,I=\varphi(G\times M))$.

$(\varphi(A_1),B_1)$ is a the subconcept of $(\varphi(A_2),B_2)$ denoted as $(\varphi(A_1),B_1)\leq(\varphi(A_2),B_2)$ if and only if $\varphi(A_1)\subseteq\varphi(A_2)$ $(\Leftrightarrow B_2\subseteq B_1)$.

Equivalently, (A_2,B_2) is the superconcept of (A_1,B_1).

Definition 5. A fuzzy concept lattice of a fuzzy formal context K with an $\alpha - Cut$ is a set C of all fuzzy concepts of K with the partial order $\leq$ with the $\alpha - Cut$ value. We noted as $\Im(C)$.

Definition 6. The similarity of a fuzzy formal concept $C_1=(\varphi(A_1),B_1)$ and its subconcept $C_2=(\varphi(A_2),B_2)$ is defined as:

$$S(C_1C_2)=\frac{|\varphi(A_1)\cap\varphi(A_2)|}{|\varphi(A_1)\cup\varphi(A_2)|} \tag{1}$$

Exemple: The corresponding fuzzy concept lattices of fuzzy context presented in table 1 are given by the following fuzzy lattices. These are illustrated in fig. 2.

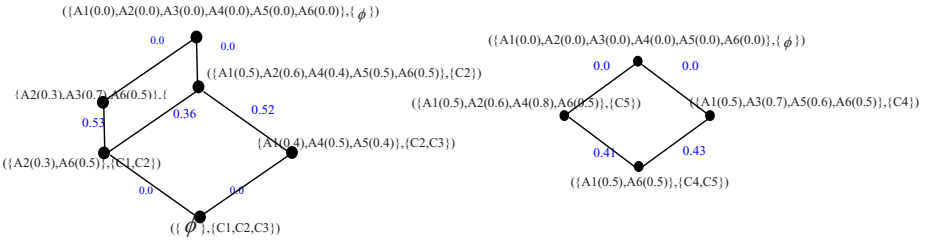

Fig. 2. The fuzzy concept lattices of the context in the Table 1

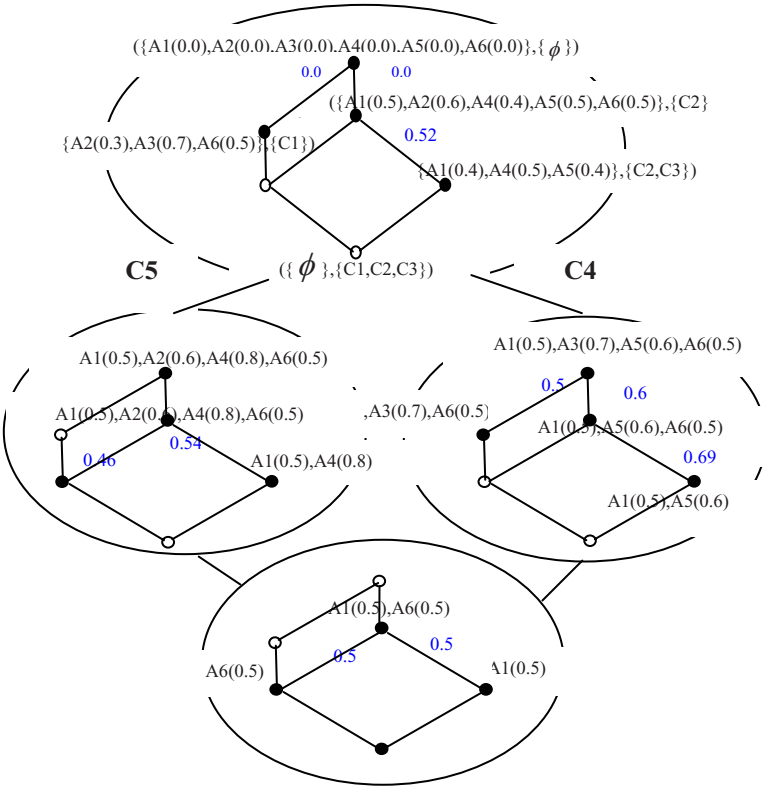

Fig. 3. A Fuzzy Nested Lattice

This very simple sorting procedure gives us for each many-valued attribute the distribution of the objects in the line diagram of the chosen fuzzy scale. Usually, we are interested in the interaction between two or more fuzzy many-valued attributes. This interaction can be visualized using the so-called fuzzy nested line diagrams. It is used for visualizing larger fuzzy concept lattices, and combining fuzzy conceptual scales on-line. Fig. 3 shows the fuzzy nested lattice constructed from Fig.2.

In this fuzzy nested line diagram, we are interested to see for each concepts of diagram represented in Fig.2 how its students are distributed in the fuzzy scale surface. We blow up each circle of fuzzy line diagram of Fig. 2 and insert the fuzzy line diagram of the surface fuzzy scale. Hence, Fig. 3 represents all pairs (c,d) of concepts c from the first and concepts d from the second fuzzy lattice. This structure is called the direct product of the two given fuzzy lattices.

From the fuzzy nested lattice, we can draw a nice usual fuzzy lattice of the same fuzzy context. This illustrated in Fig. 4.

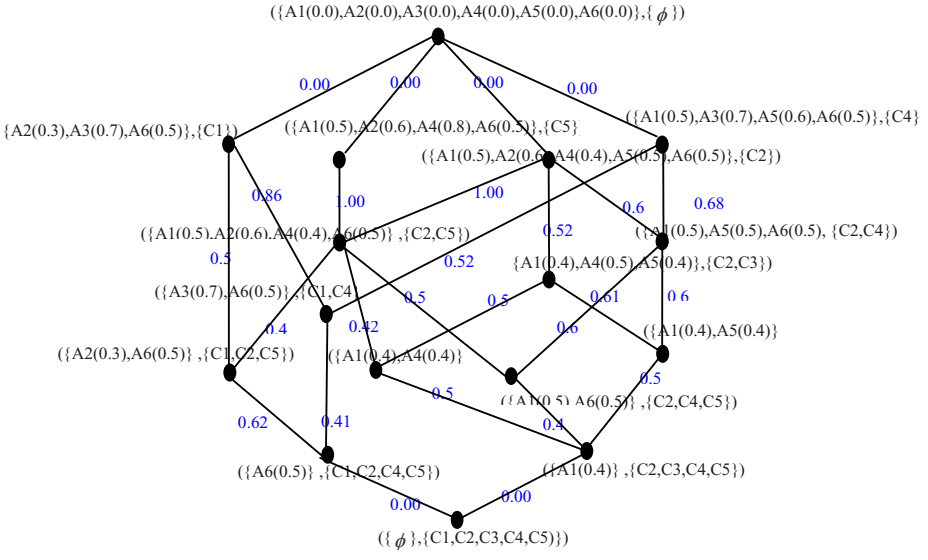

Fig. 4. A Fuzzy Clusters Lattice: FCL

3.3 Quality Evaluation

In general, the evaluation is based on a clustering validity index. The general principle of these indices consists on minimizing the compactness within a cluster and maximizing the separation between clusters.

Because they disregard lack of considering the theoretical characterization of the overlapping phenomenon, they often yield questionable results for cases involving overlapping clusters.

To cure this problem, we propose a new process of quality clustering evaluation. We give firstly an interpretation of the generated clusters and then study the quality. It consists of selection of characteristics in a given data set.

From fig. 4, we can deduce the possible overlapping between the various clusters. Let $V = \{v_j : j = 1,...,C\} \subset R^M$ a set of C clusters generated from the dataset $X = \{x_i : i = 1,...,N\} \subset R^M$

We define a distance function D as follows:

$$D: V \to \left| R^{+} \right.$$
$$\left(v_j, v_k\right) \to d$$

d is the weight of the arc connecting v_j with v_k in FCL. We note $v_j \Re v_k$ if $\exists d / D\left(v_j, v_k\right) = d$

The following properties are required:

-If $v_j \Re v_k$ and $v_k \Re v_i$ then $v_j \Re v_i$.

-If $v_j \Re v_k$ then v_j and v_k overlapped.

We study the overlapping phenomenon in the case of deducing the overlapping rate.

These properties enabled us to deduce the overlapping between different clusters. They will be used in the quality evaluation process. So, we define the separation degree and the overlapping rate. These measures form the quality index which will judge if two clusters must be merged or not.

Let $V = \left\{v_j : j = 1,..., C\right\} \subset R^M$ a set of C clusters generated from the dataset $X = \left\{x_i : i = 1,..., N\right\} \subset R^M$, $v_j \Re v_k$ and $v_k \Re v_i$ having respectively $D\left(v_j, v_k\right) = d_{jk}$ and $D(v_k, v_i) = d_{ki}$ as similarity between concepts $S\left(C_j, C_k\right)$ and $S\left(C_k, C_j\right)$. We can deduce that $D\left(v_j, v_i\right) = d_{ji} = d_{jk} + d_{ki} = S\left(C_k, C_i\right)$.

The separation Sep is given by equation 1:

$$Sep = \sum_{j=1, k \neq j}^{C} S\left(C_j, C_k\right) \tag{2}$$

In general, when Sep is large, the j^{th} and k^{th} clusters are well separated.

For example, $D\left(v_1, \left\{v_1, v_4\right\}\right) = 0.86$ and $D\left(\left\{v_1, v_4\right\}, v_4\right) = 0.52$ imply $D(v_1, v_4) = 0.86 + 0.52 = 1.38$.

We can calculate the overlapping rate, noted $Overl$, as the ratio between the number of extensions of sub-concepts, noted $Extension(Sub - Concepts)$, and the number of extensions of super-concepts, noted $Extension(Super - Concepts)$. This rate is given by equation 2.

$$Overl = \frac{\sum Extension(Sub - Concepts)}{\sum Extension(Super - Concepts)} \tag{3}$$

In general, a definition for the overlap rate implements the following principle: 1) the overlap tends to decrease $(\to 0)$ as the two components become more separated, 2) the overlap rate increases $(\to 1)$ as the two components become more strongly overlapped.

Once these requirements are met, we can evaluate the quality of clusters while basing on separation degree and the overlapping rate. We noted $Ind\_Quality$ as the quality index for a given clustering.

$$Overl = \frac{Overl}{Sep} \tag{4}$$

The clusters to be merged into one cluster are those which must maximizing the overlapping rate and minimizing the separation degree. So, a large value of $Ind\_Quality$ imply that the clusters must be merged into one.

4 Conclusion

Validity indices measure the goodness of the clustering result. A clustering is considered good if it optimizes two conflicting criteria. One of these is related to within-class scattering (the compactness), which needs to be minimized; the other to between-class scattering (the separation), which needs to be maximized. Most existing indices give good results for data sets with well separated clusters, but usually fail for complex data sets, for example, data sets with overlapping clusters.

This motivated our search for a new quality evaluation process based on Fuzzy FCA (FFCA). It consists of three steps. The first step consists of analyse the clustering results. To do this, we have proposed the fuzzy conceptual scaling notion. The second step consists of visualization. The FFCA has been proposed. It bases itself on a Fuzzy Clusters Lattice (FCL) which includes the similarity distances between different concepts in the FCL. The third step consists of evaluation the quality of generated clusters. We have defined a formal separation degree and an overlapping rate. We have defined a quality index while basing on the separation degree and the overlapping rate. The large value if this index means that the clusters must be merged into one cluster.

Future work will focus on the applicability of our quality evaluation process formal to test clustering algorithms in a more controlled way.

References

1. Menard, M., Eboueya, M.: Extreme physical information and objective function in fuzzy clustering. Fuzzy Sets and Systems 128, 285–303 (2002)
2. Bezdek, J.C.: Pattern Recognition in Handbook of Fuzzy Computation, ch. F6. IOP Publishing Ltd, Bristol (1998)
3. Sun, H.: A theory on distinguishing overlapping components in mixture models, Research Report, DMI, University of Sherbrooke, No 345 (Novenber 2003)
4. Sassi, M., Grissa Touzi, A., Ounelli, H.: Two Levels of Extensions of Validity Function Based Fuzzy Clustering. In: The 4th International Multiconference on Computer Science & Information Technology (CSIT 2006), Amman-Jordan (April 5-7, 2006)
5. Priss, U.: Formal Concept Analysis in Information Science, Annual Review of Information Science and Technology (ARIST), Preview, vol. 40 (2006)

6. Ganter, B., Wille, R.: Formal Concept Analysis: Mathematical Foundations. Springer, Heidelberg (1999)
7. Valtxhev, P., Missaoui, P., Godin, R.: Formal Concept analysis for Knowledge Discovery and Data Mining: The New Challenges. In: Eklund, P.W. (ed.) ICFCA 2004. LNCS (LNAI), vol. 2961, Springer, Heidelberg (2004)
8. Ganter, B., Wille, R.: Formal Concept Analysis: mathematical foundations. Springer, Heidelberg (1999)
9. Vogt, F., Wille, R.: TOSCANA - a graphical tool for analyzing and exploring data. In: Tamassia, R., Tollis, I.G. (eds.) Graph Drawing, pp. 193–205. Springer, Heidelberg (1994)
10. Zadeh, L.A.: Fuzzy sets. Information and Control 8, 338–353 (1965)