Towards a Suitable Reconciliation of the Findings in Collaborative Fuzzy Clustering

Rafael Falcón
Computer Science Department
Central Univ. of Las Villas (UCLV)
Santa Clara, Cuba 54830
rfalcon@uclv.edu.cu

Benoît Depaire & Koen Vanhoof
Data Analysis and Modeling
Hasselt University
3590 Diepenbeek, Belgium
benoit.depaire@uhasselt.be
koen.vanhoof@uhasselt.be

Ajith Abraham
Norwegian University of Science and Technology
Trondheim, Norway
ajith.abraham@ieee.org

Abstract

This study is concerned with the application of multi-objective particle swarm optimization (MOPSO) approaches to the framework of collaborative fuzzy clustering. In particular, the emphasis lies in determining the collaboration matrix between the data repositories. By using fitness functions both at the level of data and information granules, we can provide a more effective way of reconciling the findings between the participating data sites. A practical application of the proposed methodology to marketing research is presented.

1. Introduction

The success of Granular Computing [1] as a crucial paradigm for driving nowadays' data mining approaches can also be witnessed in clustering techniques. In recent years, several appealing clustering methods have emerged which no longer rely on data patterns to conduct the underlying optimization process but on information granules instead [2].

One of such approaches is collaborative fuzzy clustering [3][4] which can be envisioned as a collective machinery of knowledge discovery among several data sites. The key features of this innovative methodology are: (1) the exchange of information granules between the sites, as several restrictions owing to privacy or security issues are enforced over the real data, preventing them from being shared and (2) the local optimization realized at each data site becomes cognizant of the knowledge structures (clusters) uncovered at the remaining repositories and actively considers them throughout the collaborative scheme.

In order to obtain a suitable ensemble of collaborative activities leading to an overall description of the distributed data, one has to carefully regard the intensity of the collaboration $\alpha[ii, jj]$ between every pair of data repositories ii and jj. Several authors [5][6] have proposed some techniques to determine the collaboration matrix. In [6], the authors use the original similarity between the clusters at different data sites as the basis for determining the collaboration links whereas in [5], a two-step approach is presented. First, the collaboration is restricted to a subset of the available data sites, which is highly beneficial to lessen the potential communication overhead. Secondly, the very popular Particle Swarm Optimization (PSO) metaheuristic [7] is used to further optimize the collaboration matrix in terms of a single quantitative function aiming at evaluating the quality of the collaborative scheme. Yet from the experiments carried out stemmed the need for ending up with a more balanced α matrix by taking into account further criteria.

This paper presents the application of multi-objective PSO-based (MOPSO) approaches to the domain of collaborative fuzzy clustering. By considering several fitness functions expressing the quality of the collaboration realized, it is possible to know in advance how strongly the findings coming from other repositories will impact local data. The real meaning of this interaction is portrayed in terms of customer satisfaction analysis.

The most prominent features of collaborative clustering and its quantification are the subjects addressed in Sections 2 and 3, respectively whereas a concise outline of MOPSO approaches is offered in Section 4. Next, our proposal for learning the collaboration links from a new standpoint follows. Empirical results

and discussions can be found at Section 6 while conclusions finish the paper.

2. Collaborative Fuzzy Clustering

In 2002, Pedrycz [3] introduced a novel clustering algorithm, called Collaborative Fuzzy Clustering, which intended to reveal the overall structure of distributed data (i.e. data lying at different repositories) but, at the same time, complying with the restrictions preventing data sharing. It can be stated that this approach exhibits significant differences with other existing techniques under the umbrella of distributed clustering [4]. One can envision two types of collaborative clustering, namely, the horizontal mode and the vertical mode. In the first approach, each data site contains the same patterns split across different feature spaces, while in the latter approach, all data sites are described by the same feature space and different subsets of patterns are stored in separate locations. While this article focusses on the horizontal collaborative clustering scheme, the ideas presented here can be extended to the vertical fashion.

The collaborative clustering scheme first performs a local clustering analysis at the level of individual data sites, then proceeds to exchange and reconcile the knowledge structures (partition matrices and/or cluster prototypes) acting as information granules until some termination criteria is met. Any objective-function based clustering algorithm can be used as the granulation source. The augmented objective function, guiding the optimization process, can be written as

$$Q[ii] = \sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik}^2[ii]d_{ik}^2[ii] + \sum_{\substack{jj=1 \\ jj\neq ii}}^{P} \alpha[ii,jj] \sum_{k=1}^{N}\sum_{i=1}^{c}(u_{ik}[ii] - u_{ik}[jj])^2 d_{ik}[ii]. \quad (1)$$

The same notation as Pedrycz [3] is used. Here $u_{ik}[ii]$ represents the membership degree of the k^{th} pattern to the i^{th} cluster in data site ii and $d_{ik}[ii]$ representing the distance of the k^{th} pattern to the centroid of the i^{th} cluster in the ii^{th} data site. The strength of collaboration between two data sites is expressed by the collaboration link $\alpha[ii, jj]$. Finally, P, N and c stand for the number of data sites, the number of data patterns and the number of clusters respectively. The second term of Eq. 1 augments the standard FCM's objective function and realizes the communication between the data site at the level of information granules, i.e. partition matrices, instead of data patterns.

The full collaborative scheme starts with a standard FCM analysis performed at each data site, which provides the initial centroids and membership degrees. Next, the full augmented objective function is used at each data site, receiving the membership information from the other data sites from the previous step. This final step is repeated until some convergence criteria are met. For full technical details, the reader is referred to [3].

3. Quantifying the Collaboration Effect

The collaboration effect between the different data sites, can be assessed in different ways. Firstly, the original cluster prototypes (centroids) computed locally and the corresponding partition matrices will shift because of the collaboration. Secondly, the similarity between the set of clusters at at each data site will increase. Therefore, to quantify the collaboration effect, we will apply two different measures, each measuring one of the two aspects described above.

The first measure [3] compares the membership degrees of each data pattern k to each clusters i before ($u_{ik}[\mathbf{ii_{ref}}]$) and after ($u_{ik}[\mathbf{ii}]$) the collaboration. The overall impact on the partition matrices in a specific data site ii is expressed as

$$\Delta[ii] = \frac{1}{N}\sum_{k=1}^{N}\sum_{i=1}^{c} u_{ik}[\mathbf{ii}] - u_{ik}[\mathbf{ii_{ref}}]. \quad (2)$$

The average collaboration effect on the membership degrees for all data sites can then be computed as in Eq. 3. A significant variation in the membership degrees for each data site before and after the collaboration (high value of Δ) translates into a stronger collaborative impact.

$$\Delta = \frac{1}{P}\sum_{ii=1}^{P} \Delta_{ii}. \quad (3)$$

The second measure has to do with the quantification of the similarity of the final clustering outcome across the various data sites by looking at the cluster differences in terms of the memberships of patterns to clusters. Although it shows some resemblance with the similarity concept in [6], it is a completely new one. To come up with a measure of dissimilarity we redefine a cluster $C_i[ii]$ as a set of membership degrees $\{u_{1i}[ii], \cdots, u_{Ni}[ii]\}$. Now we can express the dissimilarity between cluster i from data site ii and cluster j from data site jj as follows:

$$d(C_i[ii], C_j[jj]) = \frac{1}{N} \sum_{k=1}^{N} |u_{ik}[ii] - u_{jk}[jj]|. \quad (4)$$

This dissimilarity measure will become zero, which is the lower bound, when all patterns belong to both clusters with the same degree. On the other hand, it will become equal to 1, which is the upper bound, when both clusters are crisp and don't have any pattern in common. Furthermore, this measure is also symmetric. Next, to measure the dissimilarity between the entire cluster solution of data site ii and data site jj, we compare each cluster of data site ii with each cluster of data site jj and only consider the smallest dissimilarity for each cluster (cf. Eq. 5). Note that this measure is also symmetric and equals to 0 when both cluster solutions are identical.

$$D[ii, jj] = \frac{1}{c} \sum_{i=1}^{c} \min_{j=1}^{c} [d(C_i[ii], C_j[jj])] \quad (5)$$

The final measure, which we will term as ρ, can be envisioned as the mean dissimilarity of the cluster solutions across all data sites.

$$\rho = \frac{2}{P(P-1)} \sum_{ii=1}^{P} \sum_{jj>i}^{P} D[ii, jj] \quad (6)$$

4. Multi-Objective Particle Swarm Optimization

For the sake of brevity, we assume the reader is familiar with the fundamentals of multi-objective optimization [8]. When it comes to have the PSO algorithm guide the optimization scheme, one finds that the single global best solution in one-objective optimization is replaced by a set of optimal, mutually non-dominated solutions (Pareto Front) and picking the local and global bests for every particle in the swarm is troublesome.

Most successful approaches in literature consider maintaining a bag (elite archive) where the non-dominated solutions are stored as particles move throughout the search space and from which the particle's local and global bests are selected. Pruning this archive becomes a must as its size turns unmanageable for the application unless this detrimental effect can be somehow circumvented.

We would like to consider a recent MOPSO approach for the experiments to come. Abido's novel elitist algorithm [9] maintains a population of particles and three elite archives. As soon as each particle runs into a non-dominated solution, it is added to its local best repository. The set of all Pareto-optimal solutions found by all particles in an iteration is kept in a second archive which, in turn, is dumped into a (larger) file containing the Pareto Front found so far by the algorithm. Should any of the archives exceed some predefined thresholds, it is trimmed by a clustering-driven approach. On the other hand, a particle's local and global best solutions are drawn from the local and global best archives, respectively, chosen as the individuals having the minimal distance between the two subpopulations.

Fieldsend and Singh [10] overcame the limitation imposed by pruning the elite archive by devising an efficient data structure, called "dominance tree" which allows to easily retrieve the global best for any particle. While still both local and global best repositories are in place, browsing across the latter one is very simple. Linear lists are used to store locally estimated solutions for the particles, with uniform selection of the local best.

5. Learning Collaboration Matrix with MOPSO

MOPSO approaches are a suitable way for learning the collaboration matrix in a collaborative fuzzy clustering scenario, for they can simultaneously consider several prospective solutions (particles) which, in turn, strive to achieve a balance between different criteria deemed as desirable for the ensuing granulation-driven optimization scheme. The same ideas put forward in [5] can be borrowed but this time having each particle optimize the vector $F = (-\Delta, \rho)$. The flexibility of this rationale is easily demonstrated by the fact that (1) we are not tied to any particular MOPSO implementation and (2) other quantitative functions aimed at estimating the quality of the collaborative clustering can also be incorporated to the previous vector. As an outcome, we will end up with a set of mutually non-dominated solutions (Pareto Front) from which we may draw, under some specific criteria, the individual solution that best fits our needs.

Algorithm 1 displays the breakdown of activities included in our proposal.

Achieving a suitable collaboration matrix which leads to a better reconciliation of the different knowledge structures requires several steps. The data patterns in each data site $D[ii]$ are locally optimized by any clustering vehicle (FCM, for example). Next, we might choose to limit the collaborative interaction between the repositories only to those sites for which it is expected a meaningful result (line 2, see [5]).

Algorithm 1 A MOPSO-based proposal for a balanced α determination

1: Perform local optimization in each data site $D[ii]$
2: Initialize the α matrix
3: **repeat**
4: **for** each particle x_i in the swarm **do**
5: Run the collaborative clustering
6: Evaluate the fitness function vector F
7: Select local and global best individuals
8: Update velocity v_i and position x_i
9: Update local best archive
10: **end for**
11: Update global best archive
12: **until** some termination criterion is fulfilled
13: Output the set of non-dominated solutions (Pareto Front)
14: Pick a collaboration matrix α out of the Pareto Front

Now we configure the MOPSO approach we had previously chosen and encode each particle to represent a collaboration matrix. One of the particles is initialized with the α matrix from the previous step. All of the individual prospective solutions are to conduct the full collaborative stage by using their α matrix. This stage finishes when no significant changes in some internal consistency measure are observed. Then, the quality of the resultant collaboration is estimated by means of several functions which also play the role of fitness functions and are to be optimized as a whole.

Lines 7–9 display the traditional behavior of a MOPSO algorithm and are concerned with selecting somehow the local and global best individuals for each particle in the swarm. The introduction of elitist approaches has led to increasingly good results but with an associated computational cost which might not be always overcome. Once this is done, the particles' position and velocity are updated according to the conventional rules in this sort of procedures and the local best file of the particle is updated too.

The termination criterion in line 12 often relies upon reaching a specified number of iterations. Each of them performs an update over the global best archive (Pareto Front), which is the final outcome of all MOPSO approaches. For selecting one solution out of the bunch of potential individuals, one may lean on expert criterion tied to the application domain or apply more refined heuristics to end up with the desired solution.

6. Empirical Results

6.1. Data Description

The data used for our experiments comes from a customer satisfaction survey performed in the family entertainment sector. Customers were asked to rate the performance of 54 attributes of an offering comprising four products, on scales from 1 [Low] to 10 [High]. The attribute performances were grouped into 9 dimensions and the customers also had to indicate how satisfied they were with the offering for each dimension on a scale from 1 [Low] to 10 [High]. In total, 666 respondents completed the survey entirely and were retained for our experiments. Finally, for this experiment, the original data set was divided into 9 different data sets, one for each dimension, each representing a single data site.

Table 1. Attribute Dimensions.

Attribute dimension	Number of attributes
Service	9
Product A	7
Product B	4
Quality	8
Product C	6
Product D	3
Staff	5
Prices	7
Communication	5

6.2. Experiments and Discussion

Given the context of customer satisfaction and the fact that all attributes measure performance or satisfaction from a "low-to-high" scale, we considered both 2- or 3-cluster models. Both a two-cluster analysis and a three-cluster analysis were completed by using Abido's MOPSO with the following parameters: 50 particles, 100 iterations, $c_1 = c_2 = 2$, inertia weight dynamically varied from 1.4 to 0.4.

To select the optimal number of clusters, we studied the interpretation of the two- and three-cluster models. Several cluster models from each Pareto Front were selected and their cluster prototypes for each data site were analyzed by means of a profile chart. Figures 1 and 2 show such plots for the quality dimension for a two- and three-cluster model, respectively. For the 3-cluster solution, there is only a small negligible difference between cluster 1 and 2 (cf fig. 2). This makes it very difficult to give a meaningful but different

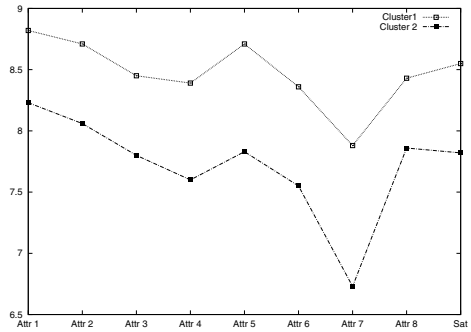

Figure 1. Quality Dimension: 2 clusters

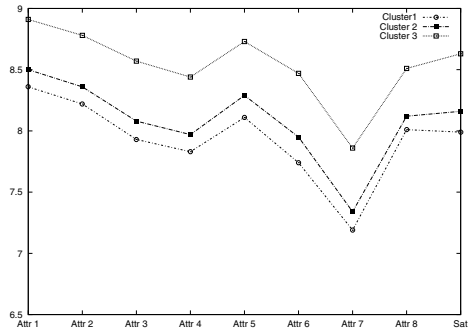

Figure 2. Quality Dimension: 3 clusters

interpretation to both clusters. The 2-cluster solution shows a similar structure, but consolidates clusters 1 and 2 from the 3-cluster solution into a single cluster (cf fig. 1). Given the small difference between these two clusters in the 3-cluster model, only little information is lost from this consolidation. The 2-cluster solution also provides a much easier interpretation of the clusters. It separates customers into "medium quality performance/satisfaction" customers and "high quality performance/satisfaction" customers. Remarkably, the same discussion also holds for the other dimensions. The 3-cluster solution always contains two clusters which are very close to each other in the feature space and the 2-cluster model consolidates these two clusters, thereby providing a more meaningful interpretation of the remaining clusters. Therefore, we shall continue to focus on the 2-cluster models.

Figure 3 shows the Pareto Front of the non-dominating 2-cluster solutions. Each particle represents a collaboration matrix. In total, 15 non-dominating solutions were found. The next problem is to select a single solution from this Pareto Front. As mentioned in section 5, this can be done by using expert knowledge or a refined heuristic. Because the number of non-dominating solutions is rather limited, we prefer to use

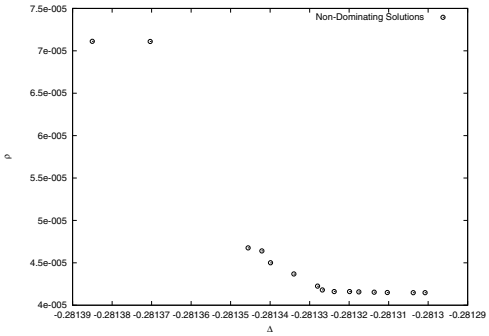

Figure 3. Pareto Front

expert knowledge to select the optimal cluster solution. From a practitioner's point of view, this approach often provides the most interesting and meaningful results, but isn't always feasible.

A first analysis of all 15 cluster solutions showed that they all revealed the same structure. The profiles of the prototypes almost never intersected, similar to fig. 1. The parallel profiles indicated that two types of customers could be revealed in each dimension: the "medium performance/satisfaction" customers and "high performance/satisfaction" customers. Thus, based on the revealed structure and the interpretation of the clusters, no solution can be preferred over the other ones.

Next, we used two different criteria to determine the best cluster solutions: i.e. the average distance between the prototypes within each dimension (data site) (discernibility) and the crispness of the clusters found. The average distance between two prototypes within one dimension is measured by taking the average absolute difference across all attributes. Next, the average distance between two prototypes were averaged across the 9 dimensions. The crispness of the clusters is quantified using the entropy criterion ([11], [12]) (cf eq. 7), with the convention that $u_{ik} \ln(u_{ik}) = 0$ if $u_{ik} = 0$. In case of crisp clusters the criterion equals to 1 and for the worst case clustering, the value of the criterion is 0.

$$I(c) = 1 - \frac{\sum_{k=1}^{N} \sum_{i=1}^{c} u_{ik} \ln(u_{ik})}{N \ln(1/k)} \qquad (7)$$

Figure 4 shows that the 8^{th} non-dominating solution provides the crispest clusters with the greatest average distance between the prototypes. Analyzing this particular cluster solution provides two interesting insights for marketeers. Firstly, it provides an easy and useful segmentation of the market into "medium

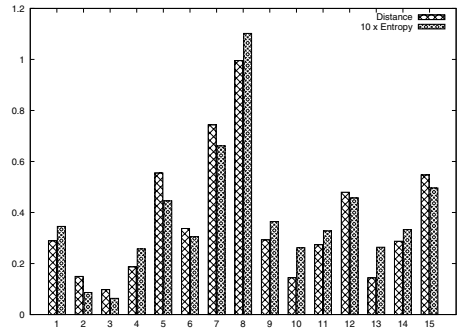

Figure 4. Discernibility and crispness of clusters for all 15 non-dominating solutions

performance/satisfaction" customers and "high performance/satisfaction" customers. Furthermore, the low value of ρ indicates that this segmentation is quasi equal in all nine dimensions, i.e. customers belong with the same membership degree to the same cluster across all nine dimensions. Secondly, the collaboration links themselves also provide interesting marketing information. Most collaboration links are very high, which indicates that much collaboration is needed to achieve clusters with the same composition across the nine dimensions. However, most interesting are the few data sites among which the collaboration link was zero, which indicates that these dimensions automatically have more or less the same cluster composition. For example, this was the case between dimension "Service" and "Product A".

7. Conclusions

We have addressed the problem of determining a suitable ensemble of collaboration links in a collaborative fuzzy clustering setting from the standpoint of a multi-objective optimization problem. The model proposed relies on any MOPSO implementation and encodes each particle as a particular collaboration matrix α.

Moreover, the fitness functions used here have to do with the evaluation of the clustering scheme and are able to describe the phenomenon of collaboration from diverse perspectives, i.e. by measuring the potential impact of the collaborative activities via cluster prototypes and partition matrices in each data site and also by looking at the cluster composition over the entire suite of data repositories after the collaboration. This leads to obtaining more balanced collaboration links which can better reflect the way in which findings are to be reconciled between the different knowledge structures.

The experimental section focused on the application of this proposal to marketing research, particularly in customer satisfaction problems. Criteria coming from the application domain allowed to select the most promising collaboration matrix so as to rule the collaboration between the nine available data repositories holding customer information.

The ideas put forward in this article can be extended in an straightforward manner to target other well-known evolutionary approaches such as Genetic Algorithms.

References

[1] W. Pedrycz, *Granular Computing: an Emerging Paradigm*, Heidelberg, Germany: Physica-Verlag GmbH, 2001.

[2] W. Pedrycz, *Knowledge-Based Clustering: from Data to Information Granules*, John Wiley & Sons, 2005.

[3] W. Pedrycz, *Collaborative Fuzzy Clustering*, Pattern Recognition Letters 23(14):1675–1686, 2002.

[4] W. Pedrycz and P. Rai, *Collaborative Fuzzy Clustering with the use of Fuzzy C-Means and its Quantification*, Fuzzy Sets and Systems, DOI 10.1016/j.fss.2007.12.030, 2008.

[5] R. Falcón, G. Jeon, R. Bello and J. Jeong, *Learning Collaboration Links in a Collaborative Fuzzy Clustering Environment*, LNCS 4827:483–495, 2007.

[6] F. Yu, J. Tang, F. Wu and Q. Sun, *Auto-weighted Horizontal Collaboration Fuzzy Clustering*, Advances in Soft Computing 40:592–600, 2007.

[7] J. Kennedy, R. Eberhart, *Particle Swarm Optimization*, IEEE Int'l Conference on Neural Networks 1942–1948, 1995.

[8] E. Zitzler, *Evolutionary Algorithms for Multiobjective Optimization: Methods and Applications*, Ph.D. Thesis, Swiss Federal Institute of Technology, Zurich, 1999.

[9] M. Abido, *Two-Level of Nondominated Solutions Approach to Multiobjective Particle Swarm Optimization*, Proceedings of the GECCO 2007, London, England, 726–733, 2007.

[10] J. Fieldsend and S. Singh, *A Multi-Objective Algorithm based upon Particle Swarm Optimization, an Efficient Data Structure and Turbulence*, Proceedings of the 2002 UK Workshop on Computational Intelligence, Birmingham, UK, pp. 37–44, 2002.

[11] G. McLachlan, D. Peel, *Finite Mixture Models*, Wiley series in probability and statistics, 2000.

[12] B. Depaire, G. Wets, K. Vanhoof, *Traffic Accident Segmentation by Means of Latent Class Clustering*, Accident Analysis and Prevention, DOI 10.1016/j.aap.2008.01.007, 2008.