Thematic Fuzzy Clusters with an Additive Spectral Approach

Susana Nascimento[1], Rui Felizardo[1], and Boris Mirkin[2,3]

[1] Department of Computer Science and Centre for Artificial Intelligence (CENTRIA), Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal
[2] Department of Computer Science, Birkbeck University of London, London, UK
[3] School of Applied Mathematics and Informatics, Higher School of Economics, Moscow, RF

Abstract. This paper introduces an additive fuzzy clustering model for similarity data as oriented towards representation and visualization of activities of research organizations in a hierarchical taxonomy of the field. We propose a one-by-one cluster extracting strategy which leads to a version of spectral clustering approach for similarity data. The derived fuzzy clustering method, FADDIS, is experimentally verified both on the research activity data and in comparison with two state-of-the-art fuzzy clustering methods. Two developed simulated data generators, affinity data of Gaussian clusters and genuine additive similarity data, are described, and comparison of the results over this data are reported.

1 Introduction

Relational data have become popular in several important application areas such as bioinformatics [25, 24, 34, 16], recommendation systems (e.g. [32]), Web mining and text analysis [22, 14, 27, 6]. Our motivation comes from our interest in mapping the activities of a research organization to a taxonomy of the field. The prime objects here are topics of the taxonomy rather than the individual members or teams in the organization, and the information is organized as an index of similarity between the topics rather than the members. In such a setting, it seems rather natural to assume an additive action of the hidden research patterns as the underlying mechanism for the generation of the similarity index. This leads us to develop a novel relational fuzzy clustering method, the Fuzzy Additive Spectral Clustering (FADDIS), by combining a model-based approach of additive clustering and the spectral clustering approach.

In spite of the fact that many relational fuzzy clustering algorithms have been developed already [2, 3, 5, 7, 9, 10, 13, 26, 33, 35], they all involve manually specified parameters such as the number of clusters or threshold of similarity without providing any guidance for choosing them. Our method does provide guidance for choosing the number of clusters. Moreover, it appears, it is quite competitive in comparison to the state of the art fuzzy clustering algorithms.

The method itself is described in a technical report [18] and briefly outlined in [20]. The main goal of this paper is to experimentally compare the FADDIS algorithm with two state-of-the-art fuzzy clustering algorithms differently

extending fuzzy c-Means to the relational data. One of these fuzzy clustering algorithms combines fuzzy c-means with a recently proposed fast-mapping technique proved superior to many other techniques, the Fast Map Fuzzy c-Means (FMFCM) [5], and the other is an extension of the c-means to dissimilarity data, the Non-Euclidean Relational Fuzzy c-Means (NERFCM) [10].

To be comprehensive in the experimentation, we developed two different cluster structure generators, each involving a controlled extent of noise. The first of them generates Gaussian entity-to-feature clusters with a different extent of intermix. The second produces genuine similarity data according to the additive fuzzy clustering model. Although the FADDIS does outperform the two other algorithms in our experiments, it also shows some unexpected behavior, which is yet to be investigated.

The rest of the paper is organized as follows. Section 2 describes the additive model and FADDIS method. Section 3 describes the experiment and its results over entity-to-feature Gaussian cluster sets. Section 4 describes the experiment and its results over genuine similarity datasets generated according to the additive fuzzy clustering data model. Section 5 illustrates application of FADDIS to the representation of thematic clusters of research activities in a hierarchic taxonomy of the field. Section 6 concludes the paper.

2 Additive Fuzzy Clustering Model and Spectral FADDIS Algorithm

The similarity, or relational, data is a matrix $W = (w_{tt'}), t, t' \in T$, of similarity indexes $w_{tt'}$, between objects t, t' from a set of objects T. Specifically, the elements of T can be leaves of a taxonomy tree such as a related hierarchical taxonomy such as Classification of Computer Subjects by ACM (ACM-CCS) [1] (see [18]). Then individual projects or members of a research organization can be represented with fuzzy membership profiles over the subjects (leaves) of the taxonomy. Given a project-to-subject profile matrix F, the similarity matrix can be defined as $W = F^T F$ so that w_{tt} is the inner product of subject columns t and t'. These subject-to-subject similarity values are assumed to be manifested expressions of some hidden patterns represented by fuzzy clusters. To develop an additive model, we formalize a relational fuzzy cluster as represented by: (i) a membership vector $\mathbf{u} = (u_t)$, $t \in T$, such that $0 \leq u_t \leq 1$ for all $t \in T$, and (ii) an intensity $\mu > 0$ that expresses the extent of significance of the pattern corresponding to the cluster. The intensity applies as a scaling factor to $\mathbf{u}$ so that it is the product $\mu\mathbf{u}$ that expresses the hidden pattern rather than its individual co-factors. Given a value of the product μu_t, to separate μ and u_t, a conventional scheme applies: the scale of the membership vector $\mathbf{u}$ is constrained on a constant level by a condition such as $\sum_t u_t = 1$ or $\sum_t u_t^2 = 1$; then the remaining factor defines the value of μ. As will be seen from formula (4), the latter normalization suits our fuzzy clustering model well and thus is accepted further on. Also, to admit a possible pre-processing transformation of the given

similarity matrix W, we denote the matrix involved in the process of clustering as $A = (a_{tt'})$.

The additive fuzzy clustering model in (1) follows that of [29, 17, 28] and involves K fuzzy clusters that reproduce the input similarities $a_{tt'}$ up to additive errors:

$$a_{tt'} = \sum_{k=1}^{K} \mu_k^2 u_{kt} u_{kt'} + e_{tt'}, \tag{1}$$

where $\mathbf{u}_k = (u_{kt})$ is the membership vector of cluster k, μ_k its intensity ($k = 1, 2, ..., K$), and $e_{tt'}$ is the residual similarity not explained by the model.

The item $\mu_k^2 u_{kt} u_{kt'}$ in (1) is the product of $\mu_k u_{kt}$ and $\mu_k u_{kt'}$ expressing the impacts of t and t', respectively, in cluster k. This value adds up to the others to form the similarity $a_{tt'}$ between topics t and t'. The value μ_k^2 summarizes the contribution of the intensity and will be referred to as the cluster's weight.

To fit the model in (1), the least-squares approach is applied, thus minimizing the sum of all $e_{tt'}^2$. Within that, the one-by-one principal component analysis strategy is attended for finding one cluster at a time by minimizing the corresponding one-cluster criterion

$$E = \sum_{t,t' \in T} (b_{tt'} - \xi u_t u_{t'})^2 \tag{2}$$

with respect to the unknown positive ξ weight and fuzzy membership vector $\mathbf{u} = (u_t)$, given similarity matrix $B = (b_{tt'})$.

In the beginning, matrix B is taken to be equal to matrix A. Each found cluster $(\mu, \mathbf{u})$ is subtracted from B, so that the residual similarity matrix applied for obtaining the next cluster is defined as $B - \mu^2 \mathbf{u}\mathbf{u}'$. In this way, A indeed is additively decomposed according to formula (1) and the number of clusters K can be determined in the process.

The optimal value of ξ at a given $\mathbf{u}$ is proven to be

$$\xi = \frac{\mathbf{u}' B \mathbf{u}}{(\mathbf{u}'\mathbf{u})^2}, \tag{3}$$

which is obviously non-negative if B is semi-positive definite.

By putting this ξ in equation (2), one arrives at $E = S(B) - \xi^2 (\mathbf{u}'\mathbf{u})^2$, where $S(B) = \sum_{t,t' \in T} b_{tt'}^2$ is the similarity data scatter.

By denoting the last item as

$$G(\mathbf{u}) = \xi^2 (\mathbf{u}'\mathbf{u})^2 = \left(\frac{\mathbf{u}' B \mathbf{u}}{\mathbf{u}'\mathbf{u}} \right)^2, \tag{4}$$

the similarity data scatter is decomposed as $S(B) = G(\mathbf{u}) + E$ where $G(\mathbf{u})$ is the part of the data scatter that is explained by cluster $(\mu, \mathbf{u})$, and E, the unexplained part. Therefore, an optimal cluster is to maximize the explained part $G(\mathbf{u})$ in (4) or its square root

$$g(\mathbf{u}) = \xi \mathbf{u}'\mathbf{u} = \frac{\mathbf{u}'B\mathbf{u}}{\mathbf{u}'\mathbf{u}}, \tag{5}$$

which is the celebrated Rayleigh quotient: its maximum value is the maximum eigenvalue of matrix B, which is reached at its corresponding eigenvector, in the unconstrained problem.

This shows that the spectral clustering approach can be applied to find a suboptimal maximizer of (5). According to this approach, one should find the maximum eigenvalue λ and corresponding normed eigenvector z for B, $[\lambda, z] = \Lambda(B)$, and take its projection to the set of admissible fuzzy membership vectors.

A number of criteria for halting the process of sequential extraction of fuzzy clusters follow from the above. The process stops if either of the conditions is true:

S1 The optimal value of ξ (3) for the spectral fuzzy cluster becomes negative.
S2 The contribution of a single extracted cluster to the data scatter becomes less than a pre-specified $\tau > 0$ threshold.
S3 The residual data scatter becomes smaller than a pre-specified $\epsilon > 0$ proportion of the original similarity data scatter.

The described one-by-one Fuzzy ADDItive-Spectral cluster extraction method is referred to as FADDIS. It combines three different approaches: additive clustering [29, 17, 28], spectral clustering [30, 23, 15, 36], and relational fuzzy clustering [9, 2, 3, 7, 5]. Since FADDIS extracts clusters one-by-one, in the order of their contribution to the data scatter, the algorithm is supposed to be oriented at cluster structures at which the clusters contribute differently the more different, the better. We refer to this supposed property of the data as the property of different contributions.

To make the cluster structure in the similarity matrix sharper, one may apply the spectral clustering approach to pre-process a raw similarity matrix W into A by using the so-called normalized Laplacian transformation which is related to the popular clustering criterion of normalized cut [15]. The normalized cut criterion can be expressed, in a relaxed form, as the minimum non-zero eigenvalue of the Laplacian matrix. To change this to the criterion of maximum eigenvalue in (5), we further transform this matrix to its pseudo-inverse matrix, which also increases the gaps between eigenvalues.

3 Experimentally Testing FADDIS on Relational Data Derived from the Entity-to-Feature Data

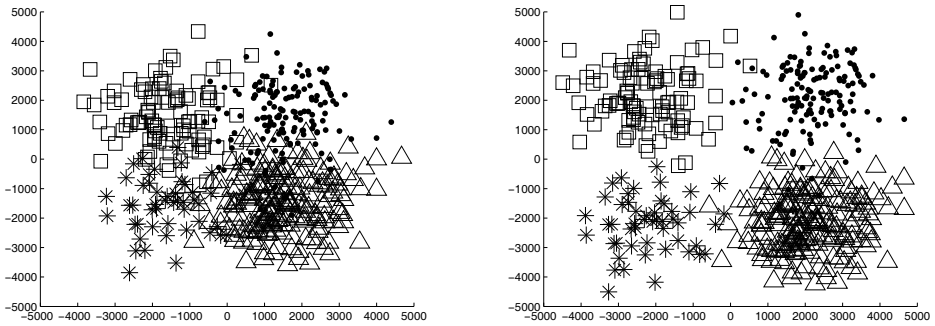
In this section, FADDIS is compared to two most effective methods for fuzzy clustering that are extensions of the popular c-means fuzzy clustering method to relational data: NERFCM [10] and FMFCM [5]. The NERFCM has been derived as an analogue to the classical c-means at the situation in which the Euclidean distance data is derived from the original entity-to-feature data. The FMFCM

also starts from the distance data to produce a number of approximating features after which the fuzzy c-means itself applies to the extracted entity-to-feature data. FADDIS applies to the affinity data derived by using the Gaussian kernel:

$$w_{tt'} = exp(-\frac{\sum_{v=1}^{V}(y_{tv} - y_{t'v})^2}{2\sigma^2}),$$

where $Y = (y_{tv})$ is a data matrix over $t \in T$ and $v = 1, 2, ..., V$, with V the number of features. The diagonal elements are set to be equal to 0: $w_{tt} = 0$ [30, 23]. The parameter σ is chosen by empirical tuning [21].

This study has been conducted with generated data based on the data generator used in [5]. Specifically, 4 clusters of data points are generated from a bivariate spherical Gaussian distribution with the standard deviation $\sigma = 950$. The centers of the clusters are defined as $c_1 = (1500, 1500)$, $c_2 = (-1500, 1500)$, $c_3 = (-1500, -1500)$, $c_4 = (1500, -1500)$, so that they are located on bisectors of the quadrants of the Cartesian plane at the same distance from the origin. The clusters have cardinalities of $50, 100, 200, 150$ data points, respectively, 500 entities altogether. In this paper, a scale parameter sn is introduced as a factor to the center of the cluster to be added to all data points, to model stretching the data points to or out of the origin. At $sn < 0$, the clusters stretch in to the origin, whereas they move out from the origin at $sn > 0$. Figure 1 illustrates the type of generated data for different values of the scale parameter sn. A data set generated at $sn = 0$ on the left, and a stretched out dataset generated at $sn = 1$ on the right.



(a) Dataset generated at $sn = 0$ (b) Dataset generated at $sn = 1$

Fig. 1: Dataset with two different scales of noise

The entity-feature generated data sets are pre-processed into dissimilarity data, to be given as input to the NERFCM algorithm, by using the matrix D of Euclidean distances between generated data points. For the FADDIS algorithm, the generated entity-to-feature data is transformed into affinity data using the

Gaussian kernel defined as $w_{ij} = exp(-d^2(y_i, y_j)/p/18)$, where d is Euclidean distance and p is the dimensionality of the data set (in our study $p = 2$). Then the Laplace Pseudo-Inverse transformation applies to transform the affinity data matrix W into the matrix A to which FADDIS algorithm is applied, which does sharpen the cluster structure in this case, as previous studies have shown [18, 8].

Ten data sets have been generated for each of the values of the scale parameter sn. The three algorithms have been run and the results have been evaluated according to the Adjusted Rand index (ARI) [12] to score the similarity between generated and computed clusterings. Also, we tested the ability of FADDIS to recover the number of clusters. In the case of FMFCM and NERFCM the number of clusters K must be prespecified; these algorithms have been applied at $K = 3, 4, 5$, after which the results have been evaluated by the extended Xie-Beni validation index [31].

Table 1 shows the means and standard deviations of the ARI index for the 10 data sets generated at each level of the scale parameter. In each row the highest ARI value is marked in boldface and (*). For the FADDIS algorithm the mode of the number of clusters retrieved by the algorithm is also presented.

The results show that FADDIS algorithm always recovers the correct number of clusters with stop condition (S2). Also, FADDIS finds the best ARI values for the data sets generated with the higher levels of cluster intermix ($sn \leq 0$). In these cases the NERFCM and FMFCM found their best partitions for a wrong number of clusters ($K = 3$). In contrast the NERFCM and FMFCM outperform the FADDIS algorithm for lower levels of cluster intermix ($sn > 0$)[1]. Yet, one should notice that the number of clusters is an input to the former algorithms.

Table 1: Bivariate Normal DG with different scale values of cluster intermix – Adjusted Rand Index (ARI) avg/std for FADDIS, NERFCM and FMFCM

	FADDIS		NERFCM			FastMap FCM		
sn	GK+Lapin	K	K = 3	K = 4	K = 5	K = 3	K = 4	K = 5
-5	**0.47/0.048***	4	**0.47/0.05***	0.44/0.05	0.37/0.035	**0.47/0.047***	0.44/0.045	0.37/0.03
0	**0.68/0.029***	4	0.66/0.034	0.64/0.058	0.53/0.032	0.66/0.035	0.61/0.096	0.54/0.013
5	0.83/0.022	4	0.76/0.018	**0.84/0.016***	0.67/0.036	0.76/0.018	**0.84/0.016***	0.67/0.031
10	0.91/0.029	4	0.82/0.015	**0.93/0.021***	0.74/0.025	0.82/0.015	**0.93/0.021***	0.75/0.029
20	0.98/0.022	4	0.86/0.008	**0.99/0.009***	0.85/0.07	0.86/0.008	**0.99/0.009***	0.82/0.067
50	**1/0***	4	0.87/0.007	**1/0***	0.87/0.075	0.87/0.007	**1/0***	0.87/0.07

4 Testing FADDIS with Genuine Similarity Data

4.1 The Fuzzy Cluster Core Data Generator

In this section, we propose a similarity data generator following the additive model (1). As usual in fuzzy clustering, we assume that each entity has one "core" cluster to which it belongs most. Therefore, the data generation process

[1] The values of the extended Xie-Beni index are concordant with the ARI values for both NERFCM and FMFCM.

starts with the generation of the "core" clusters. Then we apply the same three algorithms to the generated data.

Given the size N of an entity set I, and the number of clusters K, the proposed Fuzzy Cluster Core Data Generator (FCC DG), generates an $N \times N$ similarity data matrix G according to the underlying (FADDIS) model $W = U\Lambda U^T$, as follows:

$$G = U\Lambda U^T + \alpha E, \tag{6}$$

where:

- $N \times K$ fuzzy membership matrix U is randomly generated using a fuzzy "core" clusters generating procedure.
- Positive real valued $K \times K$ diagonal weight matrix Λ with diagonal positive values λ_k of the cluster weights equal to $\lambda_k = \mu_k^2$ is defined according to model (1). Since the vectors $\mathbf{u}_k$ in (1) are assumed normed, the weights take in the norms of the generated vectors $\mathbf{u}_k$. To test the supposed property of different contributions of the FADDIS, the weights are also made proportional to $(K - k + 1)^\beta$, for $k = 1, 2, \ldots, K$, so that the greater the $\beta > 0$, the greater the difference. Therefore, the weights are defined by $\lambda_k = (K - k + 1)^\beta * \|\mathbf{u}_k\|$.
- Elements of symmetric $N \times N$ error matrix E are independently generated from a Gaussian distribution $N(0, 1)$, and then symmetrized so that $e_{tt'} = (e_{tt'} + e_{t't})/2$.
- The value $\alpha \in [0, 1]$ is the parameter that controls the level of error introduced into the model $W = U\Lambda U^T$.

This generator builds a fuzzy cluster structure by conventionally relaxing a crisp partition. Given a crisp partition R of the entity set I, where $R = R_1, \ldots, R_K$ with non-overlapping clusters R_k, a fuzzy relaxation builds each k-th fuzzy cluster $\mathbf{u}_k$ having the corresponding crisp cluster R_k as its core in such a way that the maximum membership values u_{ik} will be at entities $i \in R_k (k = 1, \ldots, K)$ while the other components of $\mathbf{u}_k$ are close to 0.

Given the number K of core clusters covering the entire data set, I, the data generator builds each core cluster by filling it in with fuzzy membership values, such that: (a) the membership values of k-th fuzzy cluster $\mathbf{u}_k$ are very high at k-th core (e.g. $u_{ik} > 2/3$ for $i \in R_k$); and (b) the fuzzy clusters form a fuzzy partition so that $\sum_k u_{ik} = 1$ at each entity $i \in I$. After all the membership vectors $\mathbf{u}_k$ are generated, the norms of $\mathbf{u}_k$'s are computed and assigned as factors in the clusters' weights, in order to "adjust" them to the additive fuzzy clustering model. Then the final membership matrix has its membership vectors $\mathbf{u}_k$ normalized.

An example of a data set generated from the FCC DG for $K = 3$ clusters, $N = 700$ entities, and $\beta = 0.0$, visualized according to the Visual Assessment of Cluster Tendency (VAT) tool [4], is shown in Figure 2. The 3 clusters are shown in the main diagonal in dark grey, and their relative sizes can be seen. The clusters form a clear-cut structure.

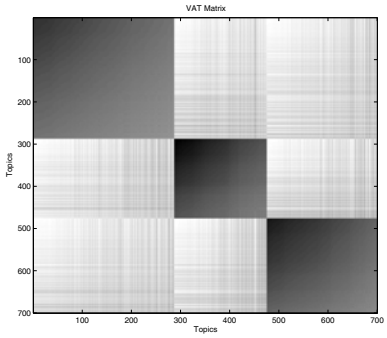


Fig. 2: VAT visualization of the cluster structure for a data set generated by the FCC DG for $K = 3$ and $N = 700$.

4.2 Setting of the experiment and its results

The main goal of this experimental study is to compare the FADDIS algorithm with FMFCM and NERFCM in recovering the cluster structures generated by the FCC DG for different levels of generated Gaussian noise.

Particular attention is given to the FADDIS algorithm, whose analysis of the clustering results are made according to the following parameters:

(i) Number of clusters retrieved by the FADDIS algorithm and corresponding stop condition achieved;
(ii) Per generated cluster k and corresponding computed cluster $\widehat{k}$, measure:
  (a) Recovery membership error (RME) of generated cluster k with membership vector $\mathbf{u}_k = [u_{ik}]$, and computed membership, $\widehat{\mathbf{u}}_k = [\widehat{u}_{ik}]$:

$$RME\,(\mathbf{u}_k) = \sum_{i=1}^{N} u_{ik}^2 \frac{|u_{ik} - \widehat{u}_{ik}|}{u_{ik}}$$

such that,

$$\sum_{i=1}^{N} u_{ik}^2 = 1$$

Notice that the RME error is an averaged relative difference weighted by u_{ik}^2, in order to normalize the error measure. The maximum value of the error is one.

  (b) Recovery intensity error (RIE) of generated and computed intensities, μ_k, and $\widehat{\mu}_k$,

$$RIE(\mu_k) = \frac{|\mu_k - \widehat{\mu}_k|}{\mu_k}.$$

(c) Percentage of the matching between generated R_k cores ($k = 1, 2, \ldots, K$) and the crisp cores retrieved from the computed partitions, after defuzzification by maximum membership;

(iii) Similarity between generated and found partitions, made according to the Adjusted Rand index (ARI) (this is to compare all the three algorithms).

The datasets have been generated in three groups corresponding to three different numbers of clusters: $K = 3, 4, 5$. The experiments were cross-combined according to the following settings: (i) Total number of entities of the data set $N = 50, 200, 400, 700$; (ii) α values of the standard deviation of noise, $\alpha = \{0, 0.05, 0.1, 0.15, 0.25, 0.5\}$. (iii) For each value of K, 10 distinct datasets had been generated for each tuple (N, α, β), resulting in a total of 720 datasets for each K value, and so a total of 2160 datasets. In the case of NERFCM, the similarity data matrix G (6) is transformed into a dissimilarity matrix D, such that, $D = max(G) - G$.

In our preliminary experiments, we observed that the ability to recover a cluster structure significantly decreases for the values of $\alpha > 0.1$. Thus, the statistics are presented for $\alpha \in \{0, 0.05, 0.1\}$ only. In the next tables, the best value in each row is marked with (*).

Table 2 shows the means/std and mode values of the recovered number of clusters by the FADDIS algorithm. For $K = 3, 4, 5$ one can see that when the β value increases from $\beta = 0.0$ to $\beta = 1.0$ the percentage of data sets for which the correct number of clusters is recovered also increases. The only exception occurs for $K = 5, N = 200$, where the best values are achieved for $\beta = 0.5$. In all the cases, the most working stop condition of the FADDIS algorithm is condition $S2$.

Table 2: FCC DG - Summary data of the percentage avg/std of correct extracted clusters and mode of the number of extracted clusters for std of added Gaussian noise=[0, 0.1] for FADDIS in best conditions for $K = \{3, 4, 5\}$

		FADDIS					
		$\beta = 0.0$		$\beta = 0.5$		$\beta = 1.0$	
	N	(%)	Mode	(%)	Mode	(%)	Mode
$K = 3$	50	50.0/0.0	3	62.5/9.6	3	85.0/5.8*	3
	200	60.0/0.0*	3	32.5/20.6	2	60.0/0.0*	3
	400	30.0/21.6	3	62.5/15.0	3	80.0/0.0*	3
	700	17.5/17.1	2	40.0/35.6	2	65.0/19.1*	3
$K = 4$	50	47.5/9.6	4	60.0/8.2	4	70.0/18.3*	4
	200	50.0/35.6	4	50.0/0.0	4	65.0/5.8*	4
	400	27.5/18.9	5	55.0/10.0	4	72.5/5.0*	4
	700	17.5/20.6	1	67.5/5.0	4	77.5/5.0*	4
$K = 5$	50	40.0/21.6	5	60.0/8.2	5	67.5/5.0*	5
	200	37.5/26.3	5	52.5/5.0*	5	40.0/8.2	5
	400	45.0/46.5	5	50.0/0.0	5	65.0/10.0*	5
	700	25.0/23.8	1	35.0/5.8	6	42.5/5.0*	5

By analysing the Recovery Membership Error (RME) and the Recovery Intensity Error (RIE) (Table 3), one can see that the minimum values are achieved for $\beta = 1.0$ for the collections of data sets with $k = 3$ and $k = 4$ clusters. For the data sets with $k = 5$ clusters the minimum values are obtained for parameter $\beta = 0.5$. Indeed, for $\beta = 1.0$ the RME and RIE mean errors are always inferior to 0.2 which is a good value (the only exception is at $K = 3$ and $N = 700$). Also, the errors almost always decrease with the increase of β, which is in accord with the expected property of different contributions of FADDIS.

Table 3: Summary Table of the RME and RIE errors' avg/std for std of added Gaussian noise=[0, 0.1] for FADDIS in best conditions for $K = \{3, 4, 5\}$

		RME			RIE		
	N	$\beta = 0.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.0$	$\beta = 0.5$	$\beta = 1.0$
$K = 3$	50	0.25/0.08	0.24/0.02	0.14/0.02*	0.14/0.03	0.15/0.01	0.08/0.01*
	200	0.28/0.08	0.54/0.12	0.15/0.01*	0.13/0.02	0.29/0.08	0.07/0.00*
	400	0.45/0.34	0.18/0.11	0.14/0.01*	0.30/0.27	0.09/0.05*	0.09/0.00*
	700	0.56/0.33	0.39/0.18	0.21/0.05*	0.35/0.24	0.25/0.14	0.13/0.05*
$K = 4$	50	0.22/0.07	0.13/0.02*	0.13/0.05*	0.11/0.01	0.07/0.01*	0.08/0.04
	200	0.44/0.35	0.12/0.01*	0.12/0.01*	0.29/0.30	0.06/0.00*	0.06/0.00*
	400	0.41/0.33	0.20/0.01	0.10/0.03*	0.28/0.29	0.10/0.00	0.05/0.01*
	700	0.59/0.36	0.13/0.05*	0.17/0.01	0.43/0.30	0.07/0.02*	0.07/0.00*
$K = 5$	50	0.28/0.12	0.14/0.02*	0.17/0.01	0.15/0.04	0.07/0.01*	0.07/0.00
	200	0.36/0.26	0.14/0.04	0.13/0.02*	0.21/0.18	0.07/0.01	0.06/0.01*
	400	0.40/0.34	0.07/0.01*	0.12/0.01	0.28/0.29	0.04/0.00*	0.05/0.01
	700	0.49/0.37	0.17/0.03*	0.18/0.01	0.35/0.33	0.10/0.01	0.06/0.00*

Table 4 presents the ARI index values for the three algorithms under consideration, FADDIS, FMFCM, and NERFCM. The highest values are marked with (*) and boldface: they always correspond to the FADDIS results. Specifically, the higher ARI values are achieved for data sets generated with $\beta = 1.0$ for the data sets with $K = 3$ and $K = 4$ clusters. For $K = 5$, the best values are achieved at $\beta = 0.5$, in contrast to the expected property of different contributions.

Complementary, and in order to compare the results obtained by the FMFCM and NERFCM algorithms the (*) mark indicates the highest ARI value between the results of these two algorithms. In almost all the cases the NERFCM outperforms FMFCM for the data sets generated with $\beta = 0.0$, which is in contrast to the case of the entity-to-feature data at which FMFCM outperforms NERFCM [5]. This illustrates the idea that the NERFCM is a genuine relational clustering algorithm whereas the FMFCM is not.

Finally, the best values for the percentages of the crisp core matching are concordant with the ARI index (not shown here).

5 Representation of Activities in a Taxonomy of the Field

As has been pointed out above, the motivation in developing the FADDIS method comes from a novel methodology of visualization of the activities of

Table 4: FCC DG - Summary Table for ARI avg/std for std of added Gaussian noise=[0, 0.1] for all algorithms in best conditions for $K = \{3, 4, 5\}$

		FADDIS			FMFCM			NERFCM		
	N	$\beta = 0.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.0$	$\beta = 0.5$	$\beta = 1.0$	$\beta = 0.0$	$\beta = 0.5$	$\beta = 1.0$
$K = 3$	50	0.88/0.14	0.84/0.21	**0.90/0.19***	0.72/0.30	0.80/0.29	0.85/0.24*	0.78/0.19	0.56/0.25	0.48/0.19
	200	0.74/0.19	0.70/0.21	**0.81/0.18***	0.45/0.35	0.56/0.31	0.62/0.32	0.69/0.23*	0.58/0.22	0.53/0.22
	400	0.87/0.10	0.87/0.10	**0.91/0.11***	0.46/0.38	0.62/0.35	0.79/0.28	0.80/0.12*	0.72/0.17	0.68/0.17
	700	0.79/0.16	0.70/0.19	**0.80/0.20***	0.36/0.36	0.51/0.38	0.58/0.32	0.70/0.21*	0.64/0.14	0.56/0.18
$K = 4$	50	0.92/0.07	0.91/0.07	**0.93/0.1***	0.65/0.28	0.77/0.18*	0.74/0.19	0.73/0.15	0.55/0.24	0.54/0.22
	200	0.92/0.09	0.91/0.09	**0.94/0.11***	0.49/0.34	0.64/0.24	0.63/0.17	0.68/0.14*	0.44/0.16	0.43/0.17
	400	0.87/0.14	0.91/0.14	**0.93/0.13***	0.42/0.35	0.59/0.28	0.7/0.23	0.72/0.13*	0.56/0.17	0.55/0.21
	700	0.84/0.15	**0.93/0.08***	0.83/0.17	0.36/0.36	0.51/0.3	0.64/0.24	0.71/0.16*	0.53/0.15	0.52/0.15
$K = 5$	50	0.92/0.08	**0.95/0.08***	0.78/0.23	0.66/0.31	0.67/0.18*	0.63/0.16	0.65/0.14	0.52/0.18	0.47/0.16
	200	0.89/0.12	**0.93/0.09***	0.83/0.17	0.48/0.33	0.63/0.22	0.6/0.16	0.64/0.16*	0.38/0.13	0.32/0.1
	400	0.87/0.22	**0.95/0.06***	0.88/0.15	0.41/0.37	0.58/0.27	0.63/0.2	0.67/0.17*	0.47/0.14	0.45/0.2
	700	0.89/0.13	**0.94/0.06***	0.84/0.13	0.36/0.36	0.52/0.29	0.61/0.19	0.66/0.16*	0.44/0.16	0.38/0.14

Table 5: A fuzzy cluster of research activities undertaken in a research centre by FADDIS

Membership value	Code	ACM-CCS Topic
0.69911	I.5.3	Clustering
0.3512	I.5.4	Applications in I.5 PATTERN RECOGNITION
0.27438	J.2	PHYSICAL SCIENCES AND ENGINEERING (Applications in)
0.1992	I.4.9	Applications in I.4 IMAGE PROCESSING AND COMPUTER VISION
0.1992	I.4.6	Segmentation
0.19721	H.5.1	Multimedia Information Systems
0.17478	H.5.2	User Interfaces
0.17478	H.5.3	Group and Organization Interfaces
0.16689	H.1.1	Systems and Information
0.16689	I.5.1	Models in I.5 PATTERN RECOGNITION
0.16513	H.1.2	User/Machine Systems
0.14453	I.5.2	Design Methodology (Classifiers)
0.13646	H.5.0	General in H.5 INFORMATION INTERFACES AND PRESENTATION
0.13646	H.0	GENERAL in H. Information Systems

a research organization such as a University department by mapping them to a related hierarchical taxonomy such as Classification of Computer Subjects by ACM (ACM-CCS) [1].

Our method generalizes the individual member/project profiles in two steps. First step finds fuzzy clusters of the taxonomy subjects according to the working of the organization. Second step maps each of the clusters to higher ranks of the taxonomy in a parsimonious way. An expository outline of this strategy, its motivations and potential benefits, made before the FADDIS has been developed, can be found in [19].

As the FADDIS found clusters are not necessarily consistent with the taxonomy, each is considered as a query set to be interpreted in the taxonomy by lifting each cluster to higher ranks of the taxonomy. The lifting is done by our recursive algorithm for minimizing a penalty function that involves "head subjects" on the higher ranks of the taxonomy together with their "gaps" and "offshoots" [20].

To illustrate the approach, Table 5 presents a fuzzy cluster obtained in our project, on the data from a survey[2] involving 16 respondents and covering 46 ACM-CCS topics, by applying the FADDIS algorithm. This cluster is then mapped to and parsimoniously generalized by the lifting method over the ACM-CCS taxonomy in terms of "head subjects" (i.e *H.-Information Systems* and *I.5-PATTERN RECOGNITION*), their "gaps" (e.g. *H.2-DATABASE MANAGEMENT*, *H.3-INFORMATION STORAGE AND RETRIEVAL*), and "offshoots" (e.g. *I.4.6- Segmentation*, *J.2- PHYSICAL SCIENCES AND ENGINEERING*). The generalized representation of the cluster resulting from the lifting method is visualized in Figure 3, pointing out its "head subjects", "gaps", and "offshoots".

[2] Survey conducted in Centre for Artificial Intelligence (CENTRIA) of Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa in 2009.

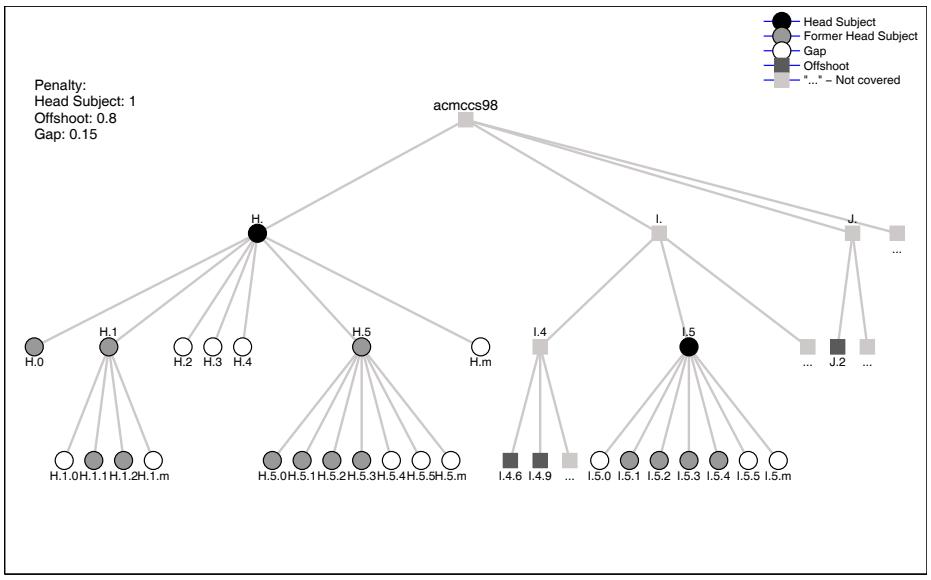


Fig. 3: Visualization of the optimal lift of the cluster in Table 1 in the ACM-CCS tree; the irrelevant tree leaves are not shown for the sake of simplicity.

6 Conclusion

The paper introduces and experimentally verifies an unconventional model of fuzzy clusters in which the products of entity membership values contribute towards similarity between the entities. This is motivated by the idea that the similarity between research topics is obtained by adding up the working of different groups on them so that the clusters according to this model can be considered thematic clusters indeed. The model leads to a spectral fuzzy clustering method FADDIS that is accompanied with a set of model-based cluster extracting stop-conditions. This paper demonstrates that FADDIS is competitive on two types of generated cluster structures. Moreover, FADDIS can be used sometimes for recovering the correct number of clusters. Yet, there are some irregularities in its working that deserve to be investigated further. One of the irregularities is the experimentally observed deviations from the property of different contributions. According to the definition of FADDIS, the more different the cluster weights in the data, that is, the greater the β at the genuine similarity data generator, the better should be the correspondence between the generated clusters and those FADDIS-computed. This is true in most cases, but sometimes it is not. We are going to address this in our future work. The other direction of further developments is applying FADDIS for visualization of activities to be captured by the analysis of web posted documents.

Acknowledgments

This work has been supported by project grant PTDC/EIA/69988/2006 from the Portuguese Foundation for Science & Technology. The partial support of the Laboratory for Analysis and Choice of Decisions in the framework of the Programme of Fundamental Studies of the National Research University Higher School of Economics, Moscow RF, to BM is acknowledged. The authors are indebted to the anonymous reviewers for multiple comments taken into account in the final version.

References

1. ACM Computing Classification System (1998), http://www.acm.org/about/class/1998 (Cited 9 Sep 2008)
2. Bezdek, J., Hathaway, R., Windham, M.: Numerical comparisons of the RFCM and AP algorithms for clustering relational data, Pattern Recognition, 24, 783-791 (1991)
3. Bezdek, J., Keller, J., Krishnapuram, R., Pal, T.: Fuzzy Models and Algorithms for Pattern Recognition and Image Processing, Kluwer Academic Publishers (1999)
4. Bezdek, J.C., Hathaway, R.J.: VAT: a tool for visual assessment of (cluster) tendency. In Procs. of the 2002 International Joint Conference on Neural Networks (IJCNN '02), 2225 - 2230 (2002)
5. Brouwer, R.: A method of relational fuzzy clustering based on producing feature vectors using FastMap. Information Sciences, 179, 3561-3582 (2009)
6. Castellano, G., Torsello, M.A.: How to Derive Fuzzy User Categories for Web Personalization. Web Personalization in Intelligent Environments, Studies in Computational Intelligence, 229/2009, Springer, 65-79 (2009)
7. Davé, R., Sen, S.: Robust fuzzy clustering of relational data, IEEE Transactions on Fuzzy Systems, 10, 713-727 (2002)
8. Felizardo, R.: A study on parallel versus sequential relational fuzzy clustering methods, Master thesis, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 212 pp. (2011)
9. Hathaway, R., Davenport, J., Bezdek, J.: Relational duals of the c-means algorithms, Pattern Recognition, 22, 205-212 (1989)
10. Hathaway, R.J., Bezdek, J.C.: NERF c-means: Non-Euclidean relational fuzzy clustering. Pattern Recognition, 27, 429-437 (1994)
11. Huang, L., Yan, D., Jordan, M.I., Taft, N.: Spectral clustering with perturbed data. In: Koller, D., Schuurmans, D., Bengio, Y., Bottou, L. (Eds.): Advances in Neural Information Processing Systems 21, Proceedings of the Twenty-Second Annual Conference on Neural Information Processing Systems (Vancouver), MIT Press, 705-712 (2009)
12. Hubert, L.J., Arabie, P.: Comparing partitions. Journal of Classification, 2, 193-218 (1985)
13. Inoue, K., Urahama, K.: Sequential fuzzy cluster extraction by a graph spectral method, Pattern Recognition Letters 20, 699-705 (1999)
14. Krishnapuram, R., Joshi, A., Nasraoui, O., Yi, L.: Low-complexity fuzzy relational clustering algorithms for Web mining. IEEE Transactions on Fuzzy Systems, 9(4), 595-607 (2001)

15. von Luxburg, U.: A tutorial on spectral clustering. Statistics and Computing, 17, 395-416 (2007)
16. Masullia, F., Mitra, S.: Natural computing methods in bioinformatics: A survey. Information Fusion, 10(3), 211-216 (2009)
17. Mirkin, B.: Additive clustering and qualitative factor analysis methods for similarity matrices. Journal of Classification, 4(1), 7-31 (1987)
18. Mirkin, B., Nascimento, S.: Analysis of Community Structure, Affinity Data and Research Activities using Additive Fuzzy Spectral Clustering. Technical Report 6, School of Computer Science, Birkbeck University of London (2009)
19. Mirkin, B., Nascimento, S., Pereira, L.M.: Cluster-lift method for mapping research activities over a concept tree. In: Koronacki, J., Wierzchon, S.T., Ras, Z.W., Kacprzyk, J. (eds.), Recent Advances in Machine Learning II, Computational Intelligence Series Vol. 263, Springer, pp. 245-258 (2010)
20. Mirkin, B., Nascimento, S., Fenner, T., Pereira, L.M.: Constructing and Mapping Fuzzy Thematic Clusters to Higher Ranks in a Taxonomy. In: Bi, Y., Williams, M.A. (eds.), 4th Intl. Conf. on Knowledge Science, Engineering & Management (KSEM 2010), Springer LNAI 6291, pp. 329-340 (2010)
21. Nadler, B., Lafon, S., Coifman, R. R., Kevrekidis, I. G.: Diffusion Maps, Spectral Clustering and Reaction Coordinates of Dynamical Systems, Applied and Computational Harmonic Analysis, (21):113-127 2006)
22. Nasraoui, O., Frigui, H.: Extracting Web User Profiles Using Relational Competitive Fuzzy Clustering. International Journal on Artificial Intelligence Tools (IJAIT), 9(4), 509-526 (2000)
23. Ng, A., Jordan, M. Weiss, Y.: On spectral clustering: analysis and an algorithm. In: Ditterich, T.G., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems, 14, MIT Press, Cambridge Ma., 849-856 (2002)
24. Pal, N.R., Aguan, K., Sharma, A., Amari, S.: Discovering biomarkers from gene expression data for predicting cancer subgroups using neural networks and relational fuzzy clustering. BMC Bioinformatics, 8(1),(5) (2007)
25. Popescu, M., Keller, J. M., Mitchell, J. A.: Fuzzy Measures on the Gene Ontology for Gene Product Similarity. Journal IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 3(3), 263-274 (2006)
26. Roubens, M.: Pattern classification problems and fuzzy sets. Fuzzy Sets and Systems 1, 239-253 (1978)
27. Runkler, T. A., Bezdek, J.C.: Web mining with relational clustering. International Journal of Approximate Reasoning, Elsevier Science, 32(2-3), 217-236 (2003)
28. Sato, M., Sato, Y., Jain, L.C.: Fuzzy Clustering Models and Applications, Physica-Verlag, Heidelberg, (1997)
29. Shepard, R.N., Arabie, P.: Additive clustering: representation of similarities as combinations of overlapping properties. Psychological Review 86, 87-123 (1979)
30. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 888-905 (2000)
31. Sledge, I.J., Bezdek, J.C., Havens, T.C., Keller, J.M.: Relational Generalizations of Cluster Validity Indices. IEEE Transactions on Fuzzy Systems 18(4), 771 - 786 (2010)
32. Suryavanshi, B.S., Shiri, N., Mudur, S.P.: An Efficient Technique for Mining Usage Profiles Using Relational Fuzzy Subtractive Clustering. Procs. of the International Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05), 23 - 29 (2005)
33. Windham, M.P.: Numerical classification of proximity data with assignment measures. Journal of Classification, 2, 157-172 (1985)

34. Xu, D., Keller, J. M., Popescu, M., Bondugula, R.: Applications of Fuzzy Logic in Bioinformatics. Imperial College Press London, UK (2008)
35. Yang, M., Shih, H.: Cluster analysis based on fuzzy relations, Fuzzy Sets and Systems 120, 197-212 (2001)
36. Zhang, S., Wang, R.-S., Zhang, X.-S.: Identification of overlapping community structure in complex networks using fuzzy c-means clustering. Physica A 374, 483-490 (2007)