A Clustering Approach using Weighted Similarity Majority Margins

Raymond Bisdorff[1], Patrick Meyer[23], and Alexandru-Liviu Olteanu[123]

[1] CSC/ILIAS, FSTC, University of Luxembourg
[2] Institut Télécom, Télécom Bretagne, UMR CNRS 3192 Lab-STICC, Technopôle Brest Iroise, CS 83818, 29238 Brest Cedex 3, France
[3] Université Européenne de Bretagne

Abstract. We propose a meta-heuristic algorithm for clustering objects that are described on multiple incommensurable attributes defined on different scale types. We make use of a bipolar-valued dual similarity-dissimilarity relation and perform the clustering process by first finding a set of cluster cores and then building a final partition by adding the objects left out to a core in a way which best fits the initial bipolar-valued similarity relation.

1 Introduction

Clustering is defined as the unsupervised process of grouping objects that are similar and separating those that are not. Unlike classification, clustering has no a priori information regarding the groups to which to assign the objects. It is widely used in many fields like artificial intelligence, information technology, image processing, biology, psychology, marketing and others. Due to the large range of applications and different requirements many clustering algorithms have been developed. Jain [14] gives a thorough presentation of many clustering methods and classifies them into partitioning [18, 17], hierarchical [11, 13, 24], density-based [2, 25], grid-based [1, 22] and model-based methods [7, 16]. New graph-based methods have also been developed in the emerging field of community detection [8, 20, 21]. Fortunato [9] covers many of the latest ones.

In this paper we present the GAMMA-S method (a Grouping Approach using weighted Majority MArgins on Similarities) for clustering objects that are described by multiple incommensurable attributes on nominal, ordinal and or cardinal scales. We draw inspiration from the bipolar outranking approach proposed by [3–5] for dealing with multiple criteria decision aid problems. As such, we assume the data is extracted in a prior stage, such that each attribute has a clear meaning and expresses a distinct viewpoint for a human agent. Also, this agent has a clear view on the importance of each attribute and what can be considered as a discriminating difference in their evaluations. For this we first characterize pairwise global similarity statements by balancing marginal similarity and dissimilarity situations observed at attribute level in order to get majority margins, i.e. a bipolar-valued similarity graph. Good maximal cliques

in this graph, with respect to a fitness measure, are chosen as cluster cores and then expanded to form a complete partition. As the enumeration of all the maximal cliques is well known to be potentially exponential [19], we develop a special meta-heuristic for dealing with the first step. The aim of our method is to achieve a partition that will minimize the differences between the original similarity relation and the relation that is implied by the clustering result.

2 Dual similarity-dissimilarity modelling

Let $X = \{x, y, z, ...\}$ denote a set of n objects. Each object $x \in X$ is described on a set $I = \{i, j, k, ...\}$ of m attributes of nominal, ordinal and/or cardinal type, where the actual evaluation x_i may be encoded without loss of generality in the real interval $[m_i, M_i]$ ($m_i < M_i \in \mathbb{R}$). The attributes may not all be of the same significance for assessing the global similarity between the objects. Therefore we assign to the attributes normalized weights $w_i \in [0, 1]$ s.t. $\sum_{i \in I} w_i = 1$, which can be given by the human agent and depend on his knowledge of the problem.

In order to characterize the *marginal pairwise similarity* and *marginal pairwise dissimilarity* relations between two alternatives x and y of X for each attribute i of I , we use the functions $s_i, d_i : X \times X \to \{-1, 0, 1\}$ defined as follows:

$$s_i(x,y) := \begin{cases} +1 & \text{, if } |x_i - y_i| \leqslant \sigma_i; \\ -1 & \text{, if } |x_i - y_i| \geqslant \delta_i; \\ 0 & \text{, otherwise.} \end{cases} \qquad d_i(x,y) := \begin{cases} -1 & \text{, if } |x_i - y_i| \leqslant \sigma_i; \\ +1 & \text{, if } |x_i - y_i| \geqslant \delta_i; \\ 0 & \text{, otherwise.} \end{cases}$$

where $0 \leq \sigma_i < \delta_i \leq M_i - m_i, \forall i \in I$ denote marginal similarity and dissimilarity discrimination thresholds. These thresholds are parameters which can be fixed by the human agent according to his a priori knowledge on the data and may be constant and/or proportional to the values taken by the objects being compared. If $s_i(x, y) = +1$ (resp. $d_i(x, y) = +1$) we conclude that x and y are similar (resp. dissimilar) on attribute i. If $s_i(x, y) = -1$ (resp. $d_i(x, y) = -1$) we conclude that x and y are not similar (resp. not dissimilar) on attribute i. When $s_i(x, y) = 0$ (resp. $d_i(x, y) = 0$) we are in doubt whether x and y are, on attribute i, to be considered similar or not similar (resp. dissimilar or not dissimilar). Missing values are also handled by giving an indeterminate $s_i(x, y) = 0$, as we cannot state anything regarding this comparison.

The *weighted similarity* and *weighted dissimilarity* relations between x and y, aggregating all marginal similarity statements and all dissimilarity statements are characterized via the functions $ws, wd : X \times X \to [-1, 1]$ defined as follows:

$$ws(x,y) := \sum_{i \in I} w_i \cdot s_i(x,y) \qquad\qquad wd(x,y) := \sum_{i \in I} w_i \cdot d_i(x,y)$$

Again, if $0 < ws(x, y) \leqslant 1$ (resp. $0 < wd(x, y) \leqslant 1$) we may assume that it is more sure than not that x is similar (resp. dissimilar) to y; if $-1 \leqslant ws(x, y) < 0$ ($-1 \leqslant wd(x, y) < 0$) we may assume that it is more sure that x is not similar (not

dissimilar) to y than the opposite; if, however, $ws(x,y) = 0$ (resp. $wd(x,y) = 0$) we are in doubt whether object x is similar (resp. dissimilar) to object y or not.

Property 1. The weighted dissimilarity is the *negation* of the weighted similarity relation: $wd = -ws$.

In some cases two objects may be similar on most of the attributes but show a very strong dissimilarity on some other attribute. In this case the objects can't be considered overall similar or dissimilar. To model this *indeterminate* situation, we define a *marginal strong dissimilarity* relation between objects x and y with the help of function $sd_i : X \times X \to \{0,1\}$ as follows:

$$sd_i(x,y) := \begin{cases} 1 & \text{, if } |x_i - y_i| \geq \delta_i^+; \\ 0 & \text{, otherwise.} \end{cases}$$

where δ_i^+ is such that $\delta_i < \delta_i^+ \leqslant M_i - m_i$ and represents a strong dissimilarity threshold. Again, this threshold is given by the human agent, in accordance with his experience concerning the underlying problem. If $sd_i(x,y) = 1$ (resp. $sd_i(x,y) = 0$) we conclude that x and y are *strongly dissimilar* (resp. not strongly dissimilar) on attribute i.

We consider that two objects x and y of X, described on a set I of attributes, are *overall similar*, denoted $(x \,\mathrm{S}\, y)$, if a weighted majority of the attributes in I validates a similarity situation between x and y and there is no marginal strong dissimilarity situation observed between x and y.

We formally characterize the *overall similarity* and *overall dissimilarity* relations by functions $s, d : X \times X \to [-1,1]$ as follows:

$$s(x,y) := \bigcirc\!\!\!\!\vee\, \big(ws(x,y), -sd_1(x,y), \cdots, -sd_m(x,y)\big)$$

$$d(x,y) := \bigcirc\!\!\!\!\vee\, \big(wd(x,y), sd_1(x,y), \cdots, sd_m(x,y)\big)$$

where, for $q \in \mathbb{N}_0$, the epistemic disjunction operator $\bigcirc\!\!\!\!\vee : [-1,1]^q \to [-1,1]$ is defined as follows:

$$\bigcirc\!\!\!\!\vee\,(p_1, p_2, \ldots, p_q) := \begin{cases} \max(p_1, p_2, \ldots, p_q) & \text{, if } p_i \geqslant 0,\ \forall i \in \{1 \ldots q\}; \\ \min(p_1, p_2 \ldots, p_q) & \text{, if } p_i \leqslant 0,\ \forall i \in \{1 \ldots q\}; \\ 0 & \text{, otherwise.} \end{cases}$$

Property 2 (Overall similarity-dissimilarity duality).
The overall dissimilarity represents the *negation* of the overall similarity:
$d = -s$.

Following from Property 2, we can now say that two objects which are not similar according to this caracterization can be called dissimilar.

For two given alternatives x and y of X, if $ws(x,y) > 0$ and no marginal strong dissimilarity has been detected, $ws(x,y) = s(x,y)$ and both alternatives are considered as overall similar. If $ws(x,y) > 0$ and a strong dissimilarity is

detected we do not state that x and y are overall similar or not, and $s(x,y) = 0$. If $ws(x,y) < 0$ and, a strong dissimilarity is observed, then x and y are certainly not overall similar and $s(x,y) = -1$. Finally, if $ws(x,y) = 0$ is observed conjointly with a strong dissimilarity, we will conclude that x and y are indeed not overall similar and $s(x,y)$ is put to -1.

We call a *Condorcet similarity graph*, denoted $G(X, s^*)$, the three-valued graph associated with the bipolar-valued similarity relation s, where X denotes the set of nodes and function $s^* : X \times X \to \{-1, 0, 1\}$ weights its set of edges as follows:

$$s^*(x,y) \ := \ \begin{cases} +1 & \text{, if } s(x,y) > 0; \\ -1 & \text{, if } s(x,y) < 0; \\ 0 & \text{, otherwise.} \end{cases}$$

3 Definition of the clusters

Ideally, a cluster would have all the objects inside it similar to each other and dissimilar from the rest. In graph theory this may be modeled by a maximal clique, however, we would also need the maximal clique to be totally disconnected from the rest of the graph, which on real data will very rarely be the case.

Therefore, in a first stage, we propose to select in the Condorcet similarity graph $G(X, s^*)$ the *best* set of maximal cliques on the $+1$ arcs (thus containing objects that are, on a majority, similar to each other), which may be considered as cluster cores. In a second stage, we expand these cores into clusters by adding objects that are well connected to them in such a way that we try to maximize the similarity arcs inside a cluster, and minimize the ones between clusters.

Let us introduce several fitness measures we will need in the algorithmic approach. Given a Condorcet outranking digraph $G(X, s^*)$ and a set $C \subseteq X$ of objects, we define, for each x of X the similarity majority margin smm towards the set C:

$$smm_C(x) \ := \ \sum_{y \in C} s^*(x,y).$$

A large positive value of $smm_C(x)$ would show that x is similar to the set C in a consistent manner. A large negative value would mean that x mostly dissimilar from C.

We define the profile of a set C by the set of all similarity majority margins for $x \in X$.

We will consider a cluster to have a strong profile is it contains strongly positive and/or negative similarity majority margins.

In order to achieve a partitioning of the entire dataset we need to detect well separated maximal cliques that correspond to local maxima of our fitness measure. To find these local maxima, we define the neighborhood of a maximal clique as all the maximal cliques that contain at least one object from it.

Let us define now the fitness an alternative x would have as part of a cluster C through function $f_C : X \to [-1, 1]$ as:

$$f_C(x) := \frac{\sum_{y \in X} s^*(x, y) \cdot smm_C(x)}{|X|^2}.$$

If x is mostly similar to C and compares to the rest of the objects in X mostly the same as the objects in C then $f_C(x)$ will be close to $+1$.

Finally we define the fitness of a partition as the outcome of the clustering method. Let f be a function that takes as argument a partition P and outputs a value inside the interval $[-1, 1]$. The fitness f of partition P is defined as:

$$f(P) := \frac{\sum_{C \in P} \sum_{x,y \in C} s(x, y) + \sum_{(C_1,C_2) \in P} \sum_{x \in C_1, y \in C_2} -s(x, y)}{\frac{|X| \cdot (|X|-1)}{2}}.$$

4 Clustering Algorithm

We structure our algorithmic approach in three steps:

1. We construct the bipolar-valued similarity relation and its associated Condorcet similarity graph.
2. We find the cluster cores.
3. We expand the cores in a greedy heuristic way.

In the second step, we may use two resolution strategies: exact enumeration of all the maximal cliques and selection of the fittest ones as potential cluster cores, or a population-based metaheuristic approach.

For the exact approach we use the Bron-Kerbosch algorithm [6], with the pivot point improvement from Koch [15]. We then evaluate the fitness of each maximal clique and compute the neighbourhood matrix from which we retrieve the maximal cliques that are the local maxima of the fitness function. As previously mentioned, the number of maximal cliques in a graph can be exponential [19], making the use of exact approaches for large or even medium clustering problems rapidly intractable.

To overcome this operational problem, we use a population-based metaheuristic close in structure to evolutionary strategies [23]. As such, the metaheuristic contains 4 steps: initialization, selection, reproduction and replacement. Each individual in the population is a maximal clique in the Condorcet similarity graph. Our aim is to discover all maximal cliques that are local maxima of our fitness measure.

In the *initialization step* we, first, iteratively generate maximal cliques that do not overlap with each other. After each object has been covered by at least one maximal clique, the rest of the population is then generated randomly.

The *selection step* has a large number of potential variations. We have opted after several tests for the rank-based roulette wheel method.

The *reproduction step* is based on a mutation operator specifically designed for maximal cliques. The maximal clique that will generate a new individual in the population is incrementally stripped with a given probability of its objects and then grown by adding other objects until the property of maximality is reached. The generated population is of equal size with the old one.

In the *replacement step*, all maximal cliques in the current population that are local maxima of the fitness measure, based however on the limited exploration of their neighborhoods that has been done at previous iterations, are kept in the new population. The rest of the individuals to be kept are selected at random.

The last step orders all the objects that were not included in a core based on their best fitness to be added to a core. The majority margins heuristic, in fact, tells us how many relations are in accordance with the decision to add the object to a particular core, therefore iteratively taking the best pair of object and core and adding that object to the core is well justified considering our goals to extract a partition that is in most accordance to the original similarity relation.

5 Results

We present some results on a few well-known datasets such as the Iris, Wines and Breast Cancer datasets from the UCI Machine Learning Repository [10].

We show the average results of our algorithm compared with the results from the classical K-means [17] and Single-Link Agglomerative Hierarchical Clustering (SL AHC) [13]. These algorithms were given the a priori knowledge regarding how many clusters the outcome should have. These results come from running every algorithm 100 times over each instance. For the first two datasets, due to their small size, we have used the exact approach of our algorithm. We have also set the thresholds to 10%, 20% and 70% of the value range on each attribute and given equal significance to all attributes.

Table 1. Average results on Jaccard Coefficient (standard deviations in brackets)

Algorithm/Dataset	Iris	Wines	Breast Cancer
K-means	0.529 (0.118)	0.461 (0.100)	0.554 (0.130)
SL AHC	0.589 (0.000)	0.336 (0.000)	0.531 (0.000)
GAMMA-S	0.603 (0.000)	0.643 (0.000)	0.560 (0.007)

We chose the *Jaccard Coefficient* [12] to measure how close the results of the clustering algorithms are to the original classes and the *Similarity Distance* to measures how close the similarity relation implied by the clustering results is to the original bipolar-valued similarity relation. This similarity relation implied by the clustering result puts $+1$ similarity values between objects inside the same cluster and -1 similarity values between objects inside different clusters. This

Table 2. Average results on Similarity Distance (standard deviations in brackets)

Algorithm/Dataset	Iris	Wines	Breast Cancer
K-means	0.710 (0.066)	0.615 (0.042)	0.600 (0.045)
SL AHC	0.707 (0.000)	0.371 (0.000)	0.590 (0.000)
GAMMA-S	0.811 (0.000)	0.693 (0.000)	0.677 (0.000)

relation would be in perfect accordance with the clustering results. This measure is in fact the weighted *Performance* measure used in Community Detection.

We notice overall that *GAMMA-S* performs better than K-means and SL AHC on both these measures. We also have to consider that we neither need to provide commensurable cardinal attributes nor an a priori number of clusters, but we take preferential information from a human agent.

6 Conclusions and Perspectives

We conclude from the testing above that our clustering method does indeed give consistent classification results, comparable with algorithms like K-means and SL AHC, however without any requirements on the data, as all kinds of attribute types can be considered. Furthermore, imprecision, uncertainties and even missing values can easily be handled by the similarity relation defined in this article. There are many improvements that could be done to increase the performance of our approach, which will be explored in the future. At the moment there is still a need to experiment with different variations in the meta-heuristic to assure a faster convergence. The final result could be further improved by means of a local search method. We also plan to explore the influence of variations of the discrimination thresholds (and the inclusion of proportional thresholds) on the outcomes.

References

1. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data, 2005.
2. M. Ankerst, M. M. Breunig, H. Kriegel, and J. Sander. Optics: ordering points to identify the clustering structure. In *International Conference on Management of Data*, pages 49–60, 1999.
3. R. Bisdorff. Logical foundation of fuzzy preferential systems with application to the electre decision aid methods. *Computers & Operations Research*, 27:673–687, 2000.
4. R. Bisdorff. Electre-like clustering from a pairwise fuzzy proximity index. *European Journal of Operational Research*, 138(2):320–331, 2002.
5. R. Bisdorff. On clustering the criteria in an outranking based decision aid approach. In *Modelling, Computation and Optimization in Information Systems and Management Sciences*, pages 409–418. Springer CCIS, 2008.

6. C. Bron and J. Kerbosch. Algorithm 457: finding all cliques of an undirected graph. *Commun. ACM*, 16(9):575–577, 1973.
7. P. Cheeseman and J. Stutz. *Bayesian Classification (AutoClass): Theory and Results*, chapter 6, pages 62–83. AAAI Press/MIT Press, 1996.
8. Nan Du, Bin Wu, Xin Pei, Bai Wang, and Liutong Xu. Community detection in large-scale social networks. In *WebKDD/SNA-KDD '07: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, pages 16–25. ACM, 2007.
9. S. Fortunato. Community detection in graphs. *Physics Reports*, 486(3-5):75 – 174, 2010.
10. A. Frank and A. Asuncion. UCI machine learning repository, 2010.
11. S. Guha, R. Rastogi, and K. Shim. CURE: an efficient clustering algorithm for large databases. In Laura Haas, Pamela Drew, Ashutosh Tiwary, and Michael Franklin, editors, *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data*, pages 73–84. ACM Press, 1998.
12. P. Jaccard. Nouvelles recherches sur la distribution florale. *Bulletin de la Societe Vaudoise de Sciences Naturelles*, 44:223–370, 1908.
13. A. Jain and R. Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.
14. A. Jain, M. Murty, and P. Flynn. Data clustering: A review. *ACM Computing Survey*, 31(3):264–323, 1999.
15. I. Koch. Enumerating all connected maximal common subgraphs in two graphs. *Theoretical Computer Science*, 250(1-2):1–30, 2001.
16. T. Kohonen. Self-organising maps. *Information Sciences*, 1995.
17. J. MacQueen. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, 1967.
18. G. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley series in probability and statistics. Wiley, 2. ed edition, 2008.
19. J. Moon and L. Moser. On cliques in graphs. *Israel Journal of Mathematics*, 3(1):23–28, 1965.
20. M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Physical Review E*, 69(2), 2004.
21. G. Palla, I. Derenyi, I. Farkas, and T. Vicsek. Uncovering the overlapping community structure of complex networks in nature and society. *Nature*, 435(7043):814–818, 2005.
22. G. Sheikholeslami, S. Chatterjee, and A. Zhang. Wavecluster: A multi-resolution clustering approach for very large spatial databases. In *VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases*, pages 428–439. Morgan Kaufmann, 1998.
23. E. Talbi. *Metaheuristics - From Design to Implementation.* Wiley, 2009.
24. T. Zhang, R. Ramakrishnan, and M. Livny. BIRCH: an efficient data clustering method for very large databases. In *SIGMOD '96: Proceedings of the 1996 ACM SIGMOD International Conference on Management of Data*, pages 103–114. ACM Press, 1996.
25. B. Zhou, D. Cheung, and B. Kao. A fast algorithm for density-based clustering in large database. In *PAKDD*, volume 1574 of *Lecture Notes in Computer Science*, pages 338–349. Springer, 1999.