Many-View Clustering - An Illustration using Multiple Dissimilarity Measures

Adán José-García
University Of Manchester
adan.jose-garcia@manchester.ac.uk

Julia Handl
University Of Manchester

Wilfrido Gómez-Flores
CINVESTAV-IPN

Mario Garza-Fabre
CINVESTAV-IPN

ABSTRACT

Multi-view problems generalize standard machine learning problems to situations in which data entities are described from multiple different perspectives, a situation that arises in many applications due to the consideration of multiple data sources or multiple metrics of dissimilarity between entities. Multi-view algorithms for data clustering offer the opportunity to fully consider and integrate this information during the clustering process, but current algorithms are often limited to the use of two views.

Here, we describe the design of an evolutionary algorithm for the problem of multi-view data clustering. The use of a many-objective evolutionary algorithm addresses limitations of previous work, as the resulting method should be capable of scaling to settings with four or more views. We evaluate the performance of our proposed algorithm for a set of traditional benchmark datasets, where multiple views are derived using distinct measures of dissimilarity. Our results demonstrate the ability of our method to effectively deal with a many-view setting, as well as the performance boost obtained from the integration of complementary measures of dissimilarity for both synthetic and real-world datasets.

KEYWORDS

Evolutionary Clustering, Multi-view Clustering, Multiple dissimilarity Measures

1 INTRODUCTION

Several applications involve multi-view data derived from (i) the consideration of multiple data sources or (ii) the application of multiple dissimilarity measures between instances. In multi-view clustering, each view captures a distinct perspective of the data and contributes with information that is necessary to fully understand the problem [2]. Multi-view algorithms for data clustering offer the opportunity to fully consider and integrate this information during the clustering process.

Here, we consider the special case of multi-view problems arising from the consideration of multiple dissimilarity measures. Existing clustering algorithms require the choice of a single measure such as the Euclidean, MED [1], or Cosine distance. The task of selecting the best dissimilarity measure for a given dataset, or that of combining multiple available measures, is typically addressed early on in the data analysis pipeline and can represent a significant challenge. One approach is to assign different weights to different measures [3, 4], but the appropriate weights are difficult to determine without prior knowledge regarding the types of structures present in the data and the reliability of the information provided by these measures. Recently, there have been some first steps toward considering multiple measures using multiobjective clustering algorithms [5, 6], but these approaches are currently limited to two-view data and do not extend to problems with many (> 3) views.

GECCO '19 Companion, July 13–17, 2019, Prague, Czech Republic

ACM ISBN 978-1-4503-6748-6/19/07…$15.00
https://doi.org/10.1145/3319619.3323365

2 THE PROPOSED ALGORITHM

We develop a methodology for many-view clustering that takes advantage of recent developments in evolutionary many-objective optimization. The framework of our algorithm kMOEA/D is outlined in Algorithm 1, and is based on the decomposition algorithm MOEA/D [8] with Tchebycheff approach, $g^{\text{te}}(\cdot)$. The algorithm requires as input: the number of subproblems (NP), a uniform spread of NP weight vectors ($\mathbf{W}$), the l dissimilarity matrices $\{\mathbf{D}_1, \ldots, \mathbf{D}_l\}$ and the termination criterion ($G_{\max}$).

Algorithm 1: General framework of kMOEA/D

Input: NP, $\mathbf{W}$, $\{\mathbf{D}_1, \ldots, \mathbf{D}_m\}$, $G_{\max}$
Output: population $\mathbf{P}$

$[\mathbf{P}, \mathbf{B}] \leftarrow$ Initialization($\mathbf{W}$)
for $g \leftarrow 1$ **to** $G_{\max}$ **do**
  for $i \leftarrow 1$ **to** NP **do**
    $\mathbf{u}_i \leftarrow$Reproduction($\mathbf{P}$,$\mathbf{B}(i)$) /* variation oper. */
    $\mathbf{C}_i \leftarrow$ Decodification($\mathbf{u}_i$, $\mathbf{w}^i$, $\{\mathbf{D}_1, \ldots, \mathbf{D}_l\}$)
    $\mathbf{f}_i \leftarrow$ Evaluation($\mathbf{C}_i$, $\{\mathbf{D}_1, \ldots, \mathbf{D}_l\}$)
    Update $\mathbf{z}^{\text{ref}}$ /* reference point */
    foreach $j \in \mathbf{B}(i)$ **do**
      if $g^{te}\left(\mathbf{u}_i \mid \mathbf{w}^j, \mathbf{z}^{ref}\right) \leq g^{te}\left(\mathbf{z}_j \mid \mathbf{w}^j, \mathbf{z}^{ref}\right)$ **then**
        $\mathbf{P}(j) = \mathbf{u}_i$, $\mathbf{Fit}(j) = \mathbf{f}_i$
      end
    end
  end
end

Initialization. kMOEA/D implements a centroid-based representation. The parent population $\mathbf{P} = \{\mathbf{z}_1, \ldots, \mathbf{z}_{NP}\}$ is randomly initialized, meanwhile, the neighborhoods $\mathbf{B}(i)$ for each i-th subproblem are selected by assigning its T closest weight vectors.

Reproduction Operators. The "DE/rand/1" mutation strategy and the binomial crossover operation are used to generate offsprings [7].

Decoding of Solutions. Let $\mathbf{m}_i = m_1, \ldots, m_K$ be the medoids derived from the i-th solution $\mathbf{z}_i$, and $\mathbf{w}^i$ be the corresponding weight vector. Then, the clustering solution $\mathbf{C}_i$ is obtained by assigning each data point $\mathbf{x}_p$ to the nearest medoid such that $\mathbf{C}_i = \text{argmin}_{j \in \mathbf{m}} d_{\text{ws}}(\mathbf{x}_j, \mathbf{x}_p)$, where $d_{\text{ws}}(\cdot, \cdot)$ is a weighted-sum distance of multiple normalized distances $\{d_1, \ldots, d_l\}$ in the direction $\mathbf{w}^i$:

$$d_{\text{ws}}(\mathbf{a}, \mathbf{b}) = w_1^i d_1(\mathbf{a}, \mathbf{b}) + \ldots + w_l^i d_l(\mathbf{a}, \mathbf{b}) \quad . \tag{1}$$

Objective Functions. The intra-cluster variance has been selected as the optimization criterion. Let $\mathbf{C}_i$ be a clustering solution and let $\mathbf{D}_j$ be a specific dissimilarity matrix, $j \in \{1, \ldots, l\}$. Then, the value for the j-th objective of the i-th subproblem is computed as $f_j(\mathbf{C}_i) = \sum_{\mathbf{c}_k \in \mathbf{C}} \sum_{\mathbf{a},\mathbf{b} \in \mathbf{c}_k} d_j(\mathbf{a}, \mathbf{b})^2$, where $d_j(\mathbf{a}, \mathbf{b})$ is the dissimilarity between $\mathbf{a}$ and $\mathbf{b}$ as defined in $\mathbf{D}_j$.

3 EXPERIMENTAL SETUP AND RESULTS

We test our algorithm for multi- and many-views settings, using various combinations of four distance functions.

Datasets. A total of 15 datasets are considered for this study: (i) nine synthetic datasets having different sizes, degrees of overlap, and clusters having different shapes; and (ii) six real-world datasets, `Iris`, `Wine`, `Breast`, `Thyroid`, `Glass` and `Ecoli`.

Parameter Settings. The settings adopted in our experiments for 2-objective instances are: $NP = 100$, $G_{\max} = 500$, the crossover rate is $Cr = 0.9$, the mutation factor is $F = 0.5$, and the neighborhood size is $T = 10$. For 3- and 4-objective instances, the population size is set to $NP = 150$ and $NP = 165$, respectively. kDE used the same settings as kMOEA/D in the case of 2-objective instances. A total of 31 independent executions were performed for each dataset. Finally, the Adjusted Rand Index (ARI) is used to assess the clustering performance. The best solution in terms of ARI was selected from the set of trade-off solutions produced by kMOEA/D.

Clustering Performance. Figure 1 summarizes the results of our experiments[1]. We observed that the single-objective algorithms obtained good results on datasets that comply with the assumptions made by the particular dissimilarity measure employed. On the other hand, kMOEA/D achieved a better performance for both synthetic and real-world datasets when simultaneously considering distinct dissimilarity measures. In general, we observed that the increase in the number of distinct dissimilarity measures systematically translates into an increase in the clustering performance.

4 CONCLUSIONS

We presented a many-objective evolutionary algorithm for clustering capable of scaling the number of views: kMOEA/D. The proposed approach outperformed some traditional single-objective clustering techniques across a diverse range of datasets, where multiple views were derived using distinct dissimilarity measures. Additionally, we investigated the performance of kMOEA/D when increasing the number of views to three and four views. Our experiments lead to the conclusion that kMOEA/D performs robustly in such a many-view setting and continues to extract value from the addition of complementary dissimilarity data. Our implementation of many-view clustering generates a set of trade-off solutions. An important challenge for future work is the automatic, unsupervised selection of the best clustering solution from these sets.

[1]Detailed results can be found in the supplementary material.

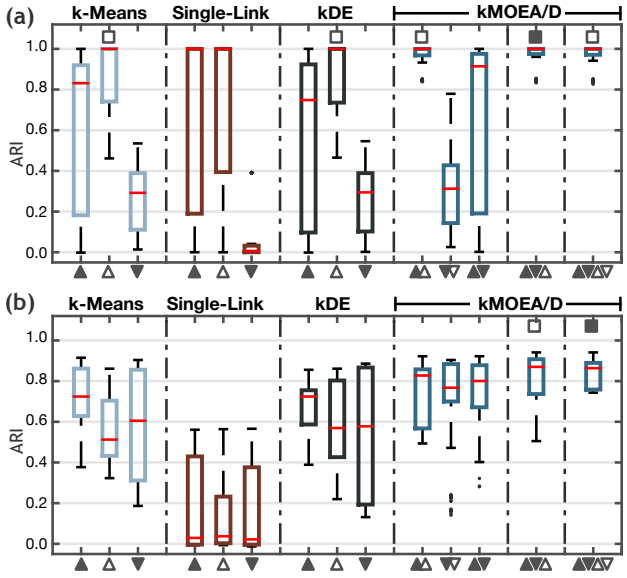

Figure 1: ARI values (best of each run) obtained for the (a) synthetic and (b) real-world datasets, using different dissimilarity measures: Euclidean (▲), MED based on Euclidean (△), Cosine (▼) and MED based on Cosine (▽). The square symbol, □, indicates no statistically significant difference between groups compared to the best one, ■.

ACKNOWLEDGMENTS

The authors are grateful to the DCSRC for a travel bursary. The first author acknowledges the support from CONACYT-Mexico through a postdoctoral fellowship.

REFERENCES

[1] Ariel E Bayá and Pablo M Granitto. 2013. How Many Clusters: A Validation Index for Arbitrary-Shaped Clusters. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 10, 2 (2013), 401–14.
[2] Guoqing Chao, Shiliang Sun, and Jinbo Bi. 2017. A Survey on Multi-View Clustering. arXiv.org, 1–17. https://doi.org/arXiv:1712.06246[cs.LG]
[3] Francisco de A.T. de Carvalho, Yves Lechevallier, and Filipe M. de Melo. 2012. Partitioning Hard Clustering Algorithms based on Multiple Dissimilarity Matrices. *Pattern Recognition* 45, 1 (2012), 447–464.
[4] J.Z. Huang, M.K. Ng, Hongqiang Rong, and Zichen Li. 2005. Automated Variable Weighting in k-means type Clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 27, 5 (2005), 657–668.
[5] Cong Liu, Qianqian Chen, Yingxia Chen, and Jie Liu. 2019. A Fast Multiobjective Fuzzy Clustering with Multimeasures Combination. *Mathematical Problems in Engineering* 2019 (2019), 1–21.
[6] Cong Liu, Jie Liu, Dunlu Peng, and Chunxue Wu. 2018. A General Multiobjective Clustering Approach Based on Multiple Distance Measures. *IEEE Access* 6 (2018), 41706–41719.
[7] Kenneth Price, Rainer Storn, and Jouni Lampinen. 2005. *Differential Evolution: A Practical Approach to Global Optimization* (natural co ed.). Springer-Verlag Berlin.
[8] Qingfu Zhang and Hui Li. 2007. MOEA/D: A Multiobjective Evolutionary Algorithm Based on Decomposition. *IEEE Transactions on Evolutionary Computation* 11, 6 (2007), 712–731.