A survey on Optimisation-based Semi-supervised Clustering Methods

Zahra Ghasemi
Department of Industrial and Systems
Engineering
Isfahan University of Technology
Isfahan, Iran
zahra.ghasemi.iut@gmail.com

Hadi Akbarzadeh Khorshidi
School of Computing and Information
Systems
The University of Melbourne
Melbourne, Australia
hadi.khorshidi@unimelb.edu.au

Uwe Aickelin
School of Computing and Information
Systems
The University of Melbourne
Melbourne, Australia
uwe.aickelin@unimelb.edu.au

Abstract— Clustering methods are developed for categorizing data points into different groups so that data points within each group have high similarities. Classic clustering algorithms are unsupervised, meaning that there is not any kind of complementary information to be utilized for attaining better clustering results. However, in some clustering problems, one may have supplementary information which can be employed for guiding the clustering process. In the presence of such information, the problem is semi-supervised clustering. In some articles, the problem of semi-supervised clustering is modeled as an optimization problem. In this research, optimization-based semi-supervised clustering papers from 2013 to 2020 are reviewed. This review is conducted based on a four-step procedure. It is attempted to explore objective functions and optimization algorithms used in these articles, as well as application domain and types of supervised information.

Keywords—*semi-supervised clustering, optimization, objective function, review*

I. INTRODUCTION

In the modern world, with enhancements in data production and storage capacities, there are a huge amount of data stored in databases. Therefore, well-equipped techniques which can categorize this data efficiently are well embraced. Data clustering is the technique that has been developed to attain this objective. Clustering algorithms attempt to group data points into distinct clusters such that data points of each cluster are most similar to each other, and at the same time are most different from the data points of other clusters [1]. Partitioning of objects in classic clustering methods is only based on the similarity of their features. These methods are named unsupervised clustering techniques, as there is no

additional information that can be employed to guide the clustering process. However, in a lot of cases, there is some sort of supervised information that can be leveraged to enhance the clustering results. For instance, cluster labels may be available for some observations, or we may know that some observations belong to the same cluster, or must belong to different clusters. Clustering algorithms that use this complementary information are named semi-supervised clustering methods [2].

Various approaches have been used for performing semi-supervised clustering. In most articles, classic clustering methods are adjusted to take advantage of supervised information. In some other articles the problem of semi-supervised clustering is modeled as an optimization problem. In these papers, one of objective functions is defined to cover supervised information.

The main goal of the current research is to review optimization-based articles in the semi-supervised clustering area of research. In this research, we aim to identify objective functions and optimization algorithms in related papers. Moreover, we try to explore application domains and types of supervised information. These aims will be presented in detail as research questions in section 2.

In order to fulfill the aims of this study, we implement our research work based on a systematic four-step method which will be presented in the second section.

In the remainder of the article, the methodology and its first three steps will be presented in section 2. Some preliminary statistics of screened papers will be presented in this section too. Then, we will attempt to answer research questions in section 3. Finally, section 4 summarizes the results and presents directions for future research.

II. RESEARCH METHODOLOGY

In this section, the four-step method for conducting the research will be presented, and its first three steps will be performed. This method contains the following steps:

- Definition of research questions: This step will determine the scope of the review.
- Extracting articles: In this step, search keywords and digital databases will be selected. By applying search keywords for selected scientific databases, initial articles will be identified.
- Screening of papers: Extracted papers in the previous step will be investigated briefly to exclude unrelated ones.
- Data extraction: Screened articles will be reviewed to answer research questions.

A. Definition of research questions

In this research, we aim to review current papers in the field of semi-supervised clustering which have modeled it as an optimization problem. The research questions we intend to answer are:

RQ1. Which optimization algorithms have been employed for semi-supervised clustering in articles?

RQ2. Which objective functions are defined in related papers?

RQ3. In which application areas, have semi-supervised clustering methods been applied?

RQ4: Which types of supervised information are utilized in articles?

B. Extracting articles

To extract related articles, search keywords and digital databases should be determined initially. The search keywords used in this study were “semi-supervised clustering”, and “optimization”. Selected digital databases for conducting the research were “IEEE Xplore”, “Springer”, and “Science Direct”. We limited our search period from 2013 to 2020. The reason for starting from 2013 is that there is a comprehensive review paper by Bair about semi-supervised clustering methods published in 2013 [2]. We will review articles after that until the most recent works. The number of extracted papers from each database is as presented in Table 1.

C. Screening of papers

After initial exploring of papers, they were investigated more to identify those articles which have dealt the semi-supervised clustering as an optimization problem. For some articles it was clear from the title of the article. For some others, it was clarified with studying abstract and introduction sections. Finally, 14 articles were identified as related articles.

TABLE I. INITIAL SEARCH RESULTS

Digital databases	URL	Number of extracted papers
Springer	http://link.springer.com/	473
IEEE Xplore	http://ieeexplore.ieee.org	103
Science Direct	http://www.sciencedirect.com	58
Total		634

Among them, just in one paper a single objective function was defined [3]. Others considered two or more objective functions. The quantity of screened papers for each year is presented in Fig. 1. As it is evident, most articles were published in 2014 and 2015, with the total number of 4 and 7 articles per year, respectively. For determining the contribution of different countries, the affiliation of the first authors is considered too. The most contributing country is India with 12 publications.

III. DATA EXTRACTION

In this section, with reviewing screened articles, research questions will be answered. We sought to explore optimization algorithms, and objective functions used in articles in our first and second research questions (**RQ1** and **RQ2**). Here we will answer these questions.

In 12 articles, the optimization algorithm is archived multi-objective simulated annealing (AMOSA) [4]. In this algorithm, the concept of pareto-dominance is incorporated into multi-objective simulated annealing (MOSA). For accepting a new solution in AMOSA, amount of dominance is considered. That is while, in previous similar researches, this acceptance is based on the number of solutions that were dominated by a new solution. In this algorithm, first, an initial archive of solutions is created. Then, one of these initial solutions is randomly selected as a current solution, and while the temperature is higher than a predetermined temperature, in each iteration this current solution is perturbed to create a new solution. After that, the domination status of the new and current solution will be checked and in each state, the current solution may be replaced by the new solution with different probabilities. In some cases, the new solution can be added to the archive set or some dominated solutions may be removed from the archive. The process of creating new solutions will be continued until reaching a predetermined number of iterations in each temperature. Then, the temperature will be reduced by a cooling rate of α and the process will be continued. Finally, when the temperature reaches the minimum threshold, the algorithm will be ended and the remained solutions in the archive will be final none-dominated solutions.

Different indices were used as objective functions in these articles for clustering n data points into K clusters as illustrated in Table 2.

Fig. 1. Number of published articles in each year

Employed indices are:

- Symmetry distance-based index (Sym-index): This index is an internal measure, which will increase when clusters are more symmetric. This index is calculated as below for K clusters:

$$Sym(K) = \left(\frac{1}{K} \times \frac{1}{\varepsilon_K}\right) \times D_K \quad (1)$$

Where $\varepsilon_K = \sum_{i=1}^K E_i$ so that $E_i = \sum_{j=1}^{n_i} d_{ps}^*(\bar{y}_j^i, \bar{z}_i)$. d_{ps}^* indicates the point symmetry-based distance between the center of the i th cluster ($\bar{z}_i$) and data point j from cluster i ($\bar{y}_j^i$) [5]. D_K is calculated as $D_K = \max_{i,j=1}^K ||\bar{z}_i - \bar{z}_j||$ that denotes the maximum distance between all pairs of cluster centers.

- Euclidean distance-based index (I-index): It is another internal index, that should be maximized. With increasing the Euclidean distance of cluster centers and decreasing the number of clusters, this index increases. I-index is computed as follows when there are K clusters [6]:

$$I(K) = \left(\frac{1}{K} \times \frac{\varepsilon_1}{\varepsilon_K} \times D_K\right)^P \quad (2)$$

ε_K is calculated as illustrated before for computing Sym-index, with this difference that here the Euclidean distance will be calculated. Similarly, D_K is the maximum Euclidean distance of cluster centers.

- Adjusted rand index (ARI): With this measure, the similarity of obtained clustering solution “C”, and true solution “T” will be determined. The method of calculating this index is as follows [6]:

$$ARI = \frac{\binom{n}{2} (e + h) - [(e + f)(e + g) + (g + h)(f + h)]}{\binom{n}{2}^2 - [(e + f)(e + g) + (g + h)(f + h)]} \quad (3)$$

Where “e” is the number of pairs of data points belonging to the same cluster in both “C” and “T” clustering solutions. In contrast, “h” denotes the number of pairs of data points assigned to different clusters in obtained and true clustering solutions. “f” and “g” are the number of pairs of data points that are clustered in the same category in one of the clustering solutions and different categories in another clustering solution.

- Minkowski index (MS-index): This index is an external measure that checks the conformity of obtained clustering solution with available supervised information of class labels. The more compliance, the lower the values of this index.

The calculation method is as follows [7]:

$$D_M(T, U) = \sqrt{\frac{n_{01} + n_{10}}{n_{11} + n_{10}}} \quad (4)$$

Where n_{11} is the total number of pairs of data points that belong to the same cluster in both true and obtained clustering solutions. n_{10} is the total number of data point pairs that must be assigned to the same cluster in reality, but in obtained clustering, they are in different clusters. In contrast, n_{01} denotes the number of data point pairs assigned in the same cluster just in obtained clustering solution.

- Fuzzy c-means index (FCM index): This index measures the distance between data points and cluster centers when data points can be assigned to more than one cluster. The calculation method is as follows [8]:

$$FCM = \sum_{i=1}^K \sum_{j=1}^n \mu_{ij}^2 ||\bar{x}_j - \bar{c}_i||^2 \quad (5)$$

In this formulation. μ_{ij} is the fuzzy membership value of j_{th} data point in i_{th} cluster.

TABLE II. OBJECTIVE FUNCTIONS OF ARTICLES BASED ON AMOSA OPTIMIZATION ALGORITHM

	Objective functions							
	Sym-index	I-index	ARI	MS-index	FCM index	XB-index	Number of features	con-index
Type Optimality	Internal max	Internal max	External max	External min	Internal min	Internal min	Internal max	Internal min
[14]	*	*						
[9]	*	*	*				*	
[16]	*	*		*				
[8]	*	*	*		*	*		
[18]	*			*		*	*	
[19]	*	*	*		*	*	*	
[10]	*		*			*	*	
[6]	*	*	*					
[7]	*	*	*					*
[15]	*	*		*				
[20]	*	*	*		*	*		
[17]	*	*		*				

- Xie and Beni index (XB-index): When clusters are compact and well-separated, this index will reduce. The calculation method is as follows [9]:

$$XB = \frac{\sum_{i=1}^K \sum_{j=1}^n \mu_{ij}^2 \|\bar{x}_j - \bar{c}_i\|^2}{n(\min_{i \neq k} \|\bar{c}_i - \bar{c}_k\|^2)} \quad (6)$$

In this index, the FCM index is divided by the multiplication of the number of data points with the minimum distance of cluster centers.

- Number of features: This objective function is aimed at increasing the number of features to compensate the effect of other internal cluster validity indexes which are biased towards lower dimensions, such as sym-index and XB-index [10].
- Connectivity-based cluster validity index (con-index): This index which measures the connectivity between clusters, is defined to secure detecting well-separated clusters with any shape, size, or convexity [7]. This index is computed as follows:

$$con = \frac{\sum_{i=1}^K \sum_{j=1}^{n_k} d_{short}(\bar{m}_i, \bar{x}_j^i)}{n \times \min_{i,j=1 \wedge i \neq j}^K d_{short}(\bar{m}_i, \bar{m}_j)} \quad (7)$$

Where $d_{short}(\bar{m}_i, \bar{x}_j^i)$ is the shortest distance along relative neighborhood graph (as illustrated in [11]) between two data point $\bar{m}_i$ and $\bar{x}_j^i$. n is the total number of data points to be clustered.

Bchir et. al, defined a joint objective function [3]. This objective function seeks to minimize the scattering of data points in clusters and the penalty for violating pairwise constraints. This objective function is as follows:

$$J = \sum_{i=1}^K \sum_{j=1}^N \sum_{k=1}^N \mu_{ij}^m \mu_{ik}^m d_{jki} + \alpha \sum_{i=1}^K \sum_{j=1}^N \sum_{k=1}^N (M - \mu_{ij}^m \mu_{ik}^m) ML(j, k) d_{jki} - \beta \sum_{i=1}^K \sum_{j=1}^N \sum_{k=1}^N (M - \mu_{ij}^m \mu_{ik}^m) CL(j, k) d_{jki} - \sum_{i=1}^N \log(\det(A_i)) \quad (8)$$

Where μ_{ij} is the fuzzy membership of data point x_j in cluster i and M is a constant. α and β denote the effect of available supervised information in form of must-link (ML) and cannot-link (CL) constraints. A_i is a matrix that will be learned for each cluster. They employed the Lagrange multiplier technique for optimizing the defined objective function in their paper.

In a recent work, Khorshidi et. al, worked on clustering patients who are injured in transport accidents [12]. They sought to find

well-separated clusters of patients which are informative regarding total medication cost as the outcome of interest. They defined two objective functions, including the cost function of the k-medians clustering, and the cross-validated error of outcome prediction. Formulation of these objective functions are as follows:

$$f(X) = \sum_{r=1}^K \mathbb{E} [I_r(Z; X) \|Z - X^r\|_1] \quad (9)$$

$$g(X) = \frac{1}{M} \sum_{m=1}^M \sqrt{\frac{\sum_{i \in F_m} (y_i - \hat{y}_i)^2}{n_m}} \quad (10)$$

$f(X)$ measures the absolute distance between data points and the nearest cluster center. For calculating $g(X)$, first regression equations are fitted to clusters, then the average error in predicting outcome variable based on obtained regression equations will be determined. In this objective function, y_i and $\hat{y}_i$ indicate real and predicted values for outcome variable respectively. M is the total number of clusters, and n_m is the number of data points in each cluster. These two objective functions are aimed at becoming minimized. For this purpose, an evolutionary computation approach based on stochastic gradient descent is employed. In this approach, first, an initial population of clustering solutions is created. Then the values of objective functions for these solutions are determined. Pareto optimal solutions will be kept for the next iteration and dominated ones will be updated based on the stochastic gradient method for k-medians clustering as proposed in [13]. This process will continue until reaching the maximum number of iterations.

In the third research question (RQ3), we sought to find application domains of screened articles. These domains include:

- Satellite image segmentation: In two papers, optimization-based semi-supervised clustering methods were employed for satellite image segmentation [14, 15]. The main objective of these articles was segmentation of image's pixels into homogenous regions. The main challenge is this fact that there are varied regions with different types of land cover. While some land covers such as bridges and roads contain smaller regions, others contain more regions.
- Magnetic resonance (MR) brain image segmentation: Two articles were identified with this application [16, 17]. In these articles, the main goal is partitioning MR brain images into non-overlapping sections for diagnosis of some brain-related disease, such as Alzheimer or Parkinson.
- Gene expression data clustering: The problem of exploring patterns in gene expression data was the most dominant application area with 5 articles [6,8,18,19,20]. It is a challenging task considering the large volume and complexity of this type of data. Gene

expression data clustering can be helpful in diagnosis of diseases.

- Identifying health service patterns: In one article, semi-supervised clustering method was used for identifying subgroups of transport accident patients who have the most similar total medication cost [12]. Findings of this research can be beneficial for health service providers to identify patients who will probably have high medication cost.

In our last research question (**RQ4**), we sought to identify different types of supervised information in reviewed articles. In 12 papers, this information was in the form of known class labels for some data points.

In some articles these class labels were extracted from data itself through clustering data points using fuzzy c-means clustering algorithm and considering 10% of data points with the highest membership values as observations with known class labels. In one article pairwise constraints were available. These constraints present that which data points must or must not belong to the same cluster. Outcome variable was the other type of supervised information in one article. In this article total cost of medication for patients was considered as supervised information.

IV. CONCLUSIONS

This paper provides a review of articles related to optimization-based semi-supervised clustering from 2013 to 2020. In these articles the problem of semi-supervised clustering is modeled as an optimization problem. This research has been conducted based on a four-step process. In this research, 634 papers have been extracted from selected databases, which have been reduced to 14 articles. With analyzing screened articles, we have sought answers of four research questions.

The findings of this research can be helpful for researchers who like to work in optimization-based semi-supervised clustering field with exploring research gaps. In most reviewed articles, class labels were used as supervised information. Developing new optimization-based semi-supervised methods employing less-considered types of supervised information can be a potential area of research. Similarly, one may wish to conduct research works for less-noticed application domains of optimization-based semi-supervised clustering. This research also illuminates employed optimization algorithms and objective functions in articles. Thus, one may work on developing new optimization-based semi-supervised clustering methods by employing new objective functions or optimization algorithms.

REFERENCES

- [1] Han, J., M. Kamber, and J. Pei, Data mining concepts and techniques third edition. The Morgan Kaufmann Series in Data Management Systems, 2011. 5(4): p. 83-124.
- [2] Bair, E., Semi-supervised clustering methods. Wiley Interdisciplinary Reviews: Computational Statistics, 2013. 5(5): p. 349-361.
- [3] Bchir, O., H. Frigui, and M.M.B. Ismail. Semi-supervised relational fuzzy clustering with local distance measure learning. in 2013 World Congress on Computer and Information Technology (WCCIT). 2013. IEEE.
- [4] Bandyopadhyay, S., et al., A simulated annealing-based multiobjective optimization algorithm: AMOSA. IEEE transactions on evolutionary computation, 2008. 12(3): p. 269-283.
- [5] Bandyopadhyay, S. and S. Saha, A point symmetry-based clustering technique for automatic evolution of clusters. IEEE Transactions on Knowledge and Data Engineering, 2008. 20(11): p. 1441-1457.
- [6] Saha, S., et al., Multi-objective semi-supervised clustering of tissue samples for cancer diagnosis. Soft Computing, 2016. 20(9): p. 3381-3392.
- [7] Alok, A.K., S. Saha, and A. Ekbal, A new semi-supervised clustering technique using multi-objective optimization. Applied Intelligence, 2015. 43(3): p. 633-661.
- [8] Alok, A.K., S. Saha, and A. Ekbal. Gene-expression data semi-supervised clustering in Multi-Objective optimization framework. in 2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI). 2014. IEEE.
- [9] Alok, A.K., S. Saha, and A. Ekbal. Feature Selection and Semi-supervised Clustering Using Multiobjective Optimization. in 2014 International Conference on Soft Computing and Machine Intelligence. 2014. IEEE.
- [10] Saha, S., et al., Feature selection and semi-supervised clustering using multiobjective optimization. SpringerPlus, 2014. 3(1): p. 1-12.
- [11] Saha, S. and S. Bandyopadhyay, Some connectivity based cluster validity indices. Applied Soft Computing, 2012. 12(5): p. 1555-1565.
- [12] Khorshidi, H.A., et al., Multi-objective semi-supervised clustering to identify health service patterns for injured patients. Health information science and systems, 2019. 7(1): p. 1-8.
- [13] Cardot, H., P. Cénac, and J.-M. Monnez, A fast and recursive algorithm for clustering large datasets with k-medians. Computational Statistics & Data Analysis, 2012. 56(6): p. 1434-1449.
- [14] Alok, A.K., S. Saha, and A. Ekbal. Pixel classification of remote sensing satellite image using semi-supervised clustering. in 2014 9th International Conference on Industrial and Information Systems (ICIIS). 2014. IEEE.
- [15] Alok, A.K., S. Saha, and A. Ekbal, Multi-objective semi-supervised clustering for automatic pixel classification from remote sensing imagery. Soft Computing, 2016. 20(12): p. 4733-4751.
- [16] Alok, A.K., S. Saha, and A. Ekbal. MR brain image segmentation using multi-objective semi-supervised clustering. in 2015 IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems (SPICES). 2015. IEEE.
- [17] Saha, S., A.K. Alok, and A. Ekbal, Brain image segmentation using semi-supervised clustering. Expert Systems with Applications, 2016. 52: p. 50-63.

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Ghasemi, Z;Khorshidi, HA;Aickelin, U

Title:

A survey on Optimisation-based Semi-supervised Clustering Methods

Date:

2021-12

Citation:

Ghasemi, Z., Khorshidi, H. A. & Aickelin, U. (2021). A survey on Optimisation-based Semi-supervised Clustering Methods. 2021 IEEE International Conference on Big Knowledge (ICBK), 00, pp.477-482. IEEE. <https://doi.org/10.1109/ickg52313.2021.00070>.

Persistent Link:

<http://hdl.handle.net/11343/297116>