A comparative study of semi-supervised clustering methods with pairwise constraint

Mostafa EL HABIB DAHO
Biomedical Engineering Laboratory
Tlemcen University, Algeria
mostafa.elhabibdaho@gmail.com

Nesma SETTOUTI
Biomedical Engineering Laboratory
Tlemcen University, Algeria
nesma.settouti@gmil.com

Salsabil LAKHDARI
Biomedical Engineering Laboratory
Tlemcen University, Algeria
salsabillakhdari@gmail.com

Amaria SAIDI
Biomedical Engineering Laboratory
Tlemcen University, Algeria
saidii.amaria@gmail.com

Mohammed El Amine BECHAR
Biomedical Engineering Laboratory
Tlemcen University, Algeria
am.bechar@gmail.com

Meryem SAIDI
Biomedical Engineering Laboratory
Tlemcen University, Algeria
miryem.saidi@gmail.com

***Abstract*—Semi-Supervised Clustering (SSC) is a largely unsupervised learning task that seeks to guide the clustering process through constraint, and combines several methods with different approaches. In this work, our interest is more focused on the semi-supervised clustering with constraint approaches, and more particularly those based on the pairwise constraint. This paper establishes a comparative study between 3 algorithms mainly: the Constrained K-means algorithm which applies constraint of comparison between pairs of objects, called COP-KMEANS, the Semi-supervised kernel clustering with relative distance Algorithm (SKLR) and the Semi-supervised Kernel Mean Shift clustering Algorithm (SKMS). Experimental results indicate that the semi-supervised kernel Mean Shift clustering method can generally outperform the other semi-supervised methods. The experimental study shows that the use of constraint can improve performance especially when the number of available labeled examples is insufficient to build a decision model.**

***Index Terms*—Semi-Supervised Clustering, learning constraint, pairwise constraint, COP K-means, Kernel Mean Shift Clustering, Kernel clustering with Relative distance.**

I. Introduction

In the context of semi-supervised learning, the techniques are grouped together taking into account all the partially labeled samples. We, therefore, note the learning sample S composed of a supervised sample S_{sup} and an unsupervised S_{unsup}:

$$S = S_{sup} \cup S_{unsup}.$$

In recent years, the unsupervised learning or clustering has been addressed by various researchers; most of these works are interested in integrating constraint into these methods. These constraint can be generated from prior knowledge of the data [1], or from a subset of labeled data [2]. Taking this knowledge into account in a clustering process, represents a new field of study in machine learning known as constraint classification [3]. In addition, incorporating prior knowledge into clustering processes has gained momentum in a number of real-world applications such as face recognition via surveillance cameras [4], the refinement of GPS maps [5] and landscape detection in hyper-spectral data [6] [7].

Moreover, during these last ten years, a lot of works on the integration of constraint in the unsupervised learning methods are seen to be of great interest [8]–[10]. They represent a new field of study in semi-supervised learning recognized as semi-supervised learning with constraint.

In our work, we are interested in the study of the semi-supervised clustering approaches with constraint classification of medical data. To do this, we carry out a comparative study of three recent semi-supervised constraints techniques: cop-kmeans [5], the semi-supervised Mean Shift clustering [11] and the semi-supervised kernel clustering with relative distance [12]. The objective is to discuss and analyze moreover the influence of the pairwise constraint (Must-Link and Cannot-Link) on the performances of clustering by carrying out experiments with different percentages of labeled examples.

The study is conducted on 6 medical data sets selected from the UCI repository [13], and the paper is structured as follows: we briefly present the main types of semi-supervised clustering constraint in Section 2. Subsequently, section 3 will summarize the state of the art of the different types of constraint and methods used in semi-supervised clustering. Section 4 details the principles of the three semi-supervised clustering techniques chosen for the comparative study. In section 5, we discuss the experimental results obtained by the comparative study on different medical datasets. Finally, a general conclusion and prospects of the work come to close this paper.

II. Constraint clustering

Constraint clustering is an important task in the data mining process. It allows modeling more finely the clustering task by integrating constraint of users. See figure Fig.1.

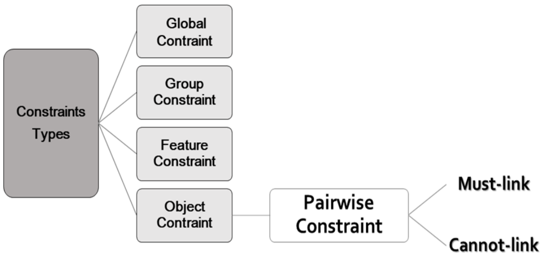

Fig. 1. The different types of constraint

Several types of constraint can be considered [14]; they can be related to clusters, such as their diameter or size, or can be related to pairs of objects that must be in the same class or not. A constraint can also depend on the type of information that is manipulated, the latter can take several forms.

A. Global constraint

If the information is of structural form on the data, in this case there are 4 ways to proceed :

- **Adding position attributes:** This method makes it possible to add a notion of spatial location in a set of data, this is done in a way to add positional attributes for each object.
- **Duplication of neighbors:** This duplication technique is based on the notion of neighborhood in terms of distance d_{ij} in D dimensions (with $i \neq j$). It makes it possible to increase the size of the vector attribute of an object with the addition of one or more sets of attributes according to the number of neighbors considered [15].
- **Changing the distance calculation:** Unlike the previous methods of modifying the set of data (and more particularly, the vector attribute of each object by the addition of new attributes), in a less direct way certain methods make it possible to integrate constraint by incorporating spatial information, this by changing the way of calculating the distance between two objects.
- **Modification of the objective function:** This last category of global constraint methods, and more particularly neighborhood information, modifies a criterion to be optimized by the objective function of any algorithm, using an optimization procedure.

B. Group constraint

In machine learning, available knowledge can also be provided from groups information of objects. It can be defined as requirements on the overall shape, orientation or other characteristics of the groups. The minimum or maximum capacity of these seems to be the most used characteristic in the literature.

- **Minimum capacity constraint :** Methods that use constraint on groups of objects, which are expanded to avoid solutions containing empty groups. This approach imposes constraint on the structure of groups. As a result, it is possible to specify a minimum number of objects for each group.
- **Maximum capacity constraint :** The use of this type of constraint is most often applied by non-supervised learning algorithms of the hierarchical grouping type. Consequently, from the created hierarchy, it is possible to select groups of adapted objects making it possible to enforce the defined constraint.

C. Features constraint :

Prior knowledge can be interpreted as information dependent on the characteristics of objects. These constraint make it possible to orient the classification of objects according to their values for a given feature.

D. Objects constraint :

In 1984, Bejar and Cortes [16] developed a constrained hierarchical classification algorithm incorporating feature constraint. Object constraint define boundaries on individual pairs of objects. This type of prior knowledge about data is usually provided in three forms:

- **Partial labeling :** The labeling of a higher dimension set of objects is represented by a complex and expensive function in computing time which makes this task often impossible. As a solution, it is possible to label a subset containing only a few objects.
- **Feedback :** An iterative approach has been adopted by the interactive classification systems, this system produces a partition of data then evaluates and validates it by an expert, if all the data is of important dimensions, we can find a difficulty during the validation of the results, so an expert can clearly indicate the partitioning errors (system-induced) after we can use this information in the next iteration.
- **Relationship between pairwise objects :** The constraint produce indications on the desired partition and implement these indications in clustering algorithms to increase their performance [1]. Let $X = x_1, ..., x_n$ be the set of observations that must be grouped into K classes, and denoted by $u_1, ..., u_K$. For each pair of observations x_i, x_j in X, we denote the distance between them by $d(x_i, x_j)$. This type of constraint simply certifies that two objects are:

  a) Must-Link (ML): which forces two observations x_i and x_j to be in the same class.

Where for two instances of data x_i and x_j in the data set, $x_i, x_j \in X(1 \leq i, j \leq n)$, if x_i and x_j satisfy the Must-Link constraint, then after completing the clustering, x_i and x_j satisfy $x_i \in Cm \wedge x_j \in Cm, Cm \in \prod, 1 \leq m \leq k$, otherwise the cluster fails. The constraint can be described as x_i ML x_j.

b) Cannot-Link (CL): If two observations x_i and x_j are in two different classes we can define these two objects by Cannot-Link (CL). Where for two instances of data x_i and x_j in the data set, $x_i, x_j \in X(1 \leq i, j \leq n)$, if x_i and x_j satisfies the Cannot-Link constraint, after completing the clustering, x_i and x_j satisfy $x_i \in Cm \wedge x_j \in Cn, Cm, Cn \in \prod, 1 \leq m, n \leq k, m \neq n$, otherwise the cluster fails. The constraint can be described as x_i CL x_j.

In this paper, we are interested in the comparative study of semi-supervised learning algorithms that are based on the pairwise constraint and are provided as two types of constraint: Must-Link (ML) and Cannot-Link (CL).

III. Related work on Semi-Supervised Clustering (SSC)

The emergence of Semi-Supervised Clustering (SSC) techniques appeared by the extension of unsupervised methods (clustering) in semi-supervised learning. The adaptation was done in two groups of unsupervised learning methods : feature-based and graph-based approaches.

A. Feature-based approach:

where each data point has a representation in terms of a feature, a vector or a structured representation as a sequence, time series, or graphic. e.g. k-means and mixture of Gaussian....

B. Graph-based approach:

where a graphical similarity between data points is granted, e.g. spectral clustering methods.

Indeed, the literature shows that a large number of works have focused on the application of these two types of semi-supervision (Fig.2) :

- **Pointwise :** where the cluster label of a small number of points are available to guide clustering two by two. [17]
- **Pairwise :** where ”must-link” and ”cannot-link” the constraint between certain pairs of points are available.

Over the past decade [18], feature-based methods have been widely studied for proper generalization, algorithms such as k-means and its variants have been one of the most successful adaptation that has accurately drawn part of the semi-supervision. These methods have been generalized to incorporate the SSC context learning metric and the parameter and inference estimation into a graphical model [2] [19] [20].

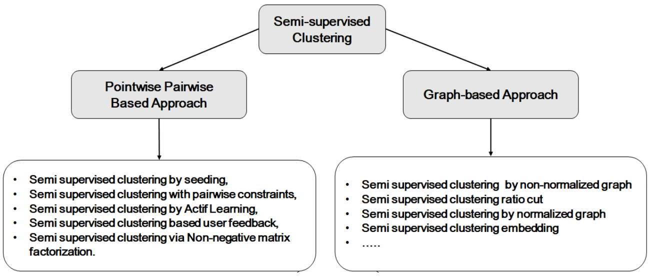
Semi-supervised Clustering
Pointwise Pairwise Based Approach
Graph-based Approach
• Semi supervised clustering by seeding,
• Semi supervised clustering with pairwise constraints,
• Semi supervised clustering by Actif Learning,
• Semi supervised clustering based user feedback,
• Semi supervised clustering via Non-negative matrix factorization.
• Semi supervised clustering by non-normalized graph
• Semi supervised clustering ratio cut
• Semi supervised clustering by normalized graph
• Semi supervised clustering embedding
• .....

Fig. 2. The Semi-Supervised Clustering (SSC) Approaches

On the other hand, there are several approaches to SSC based on graphical representation and clustering methods based on graphs [21], [22]. The literature on SSC has also exploited the graph track centered mainly on points to explore in the semi-supervised context. The methods of spectral clustering are widely used for unsupervised learning with the graphical presentation [19] [3] [23] and at the same time, can be considered as a solution of graph-cut space problems.

IV. Methods

The aim of this work is to discuss and analyze moreover the influence of the pairwise constraint (must-link and cannot-link) on the performances of clustering by carrying out experiments with different percentages of labeled examples. In this direction, we carry out a comparative study of three recent semi-supervised clustering learning techniques with pairwise constraint : cop-kmeans [5], the semi-supervised Mean Shift clustering [11] and the semi-supervised kernel clustering with relative distance [12].

A. Constrained K-means Algorithm (COP K-means)

The goal of clustering analysis methods is to share a set of data into homogeneous subgroups. In literature, we notice that most existing semi-supervised clustering methods are modified versions of k-means [24]. The COP K-means algorithm of Wagstaff et al. [5] is the most advanced of these versions, by integrating the constraint, it has been demonstrated that this approach makes it possible to guide and improve clustering, and thus improves the performances of the algorithms. As part of the semi-supervised clustering, there are many ways to constrain data. COP K-means consider particularly two types of possible constraint between observations: if the two observations are in the same cluster we have the type ”must -link”; otherwise they are in a different cluster so this constraint is ”cannot-link”. The algorithm takes a set of data (D), a set of must-link $constraint(Con_{=})$, and a set of constraint not-link $(Con_{\neq})$. It returns a partition of the instances in D that satisfies all the specified constraint.

B. The Semi-supervised Kernel Mean Shift clustering Algorithm (SKMS)

The goal of this method is to integrate supervision into the mean shift clustering method [25] which uses

Algorithm 1 :COP K-means

```
Input:
$D$ : a set of data
$(Con_{=})$:a set of Must-Link constraint
$(Con_{\neq})$:a set of Cannot-Link constraint

1: Select randomly : $Kpoints$ : initial cluster centers.
2: Each point $d_i \in D$ is assigned to its nearest cluster while
   ensuring no constraint $(Con_{=})$ and $(Con_{\neq})$ is broken.
3: Update each center cluster to be the means of its constituent items
4: Repeat (2) and (3) until convergence.
   Return $(Con_{=}), (Con_{\neq})$
```

only pairwise constraint to guide the clustering procedure.

Mean shift clustering: proposed by Cheng et al. [25] is a powerful non-parametric popular mode search technique that does not require prior knowledge of the number of clusters and does not limit the shape of clusters.

- A popular mode that iteratively locates modes in data by maximizing kernel density estimate (KDE).
- The non-parametric nature of mean shift makes it a powerful tool for discovering arbitrarily shaped clusters present in the data. In addition, the number of clusters is automatically determined by the number of discovered modes.

The semi-supervised kernel Mean Shift clustering (SKMS) algorithm of Anand et al. [11] generalizes the linear projection operation to a linear transformation of the kernel space which will make it possible to scale the distance between the constraint points. With this transformation, the must-link points are moved closer together, while the can-not points can be moved further. This transformation is performed as shown in algorithm 2. For each cluster in the dataset, a small amount of labeled data is used to generate the even constraint (must-link and cannot-link).

C. The Semi-supervised kernel clustering with relative distance Algorithm (SKLR)

This algorithm proposed by Amid et al. [12] is largely inspired by the SKMS algorithm. The main contribution is to extend the SKMS algorithm to handle relative distance comparisons. This, to consider the problem of clustering a set of data into k groups subject to an additional set of constraint on comparisons of relative distance between data elements. The additional constraint are intended to preselect lateral information that is not expressed directly in the feature vectors.
Relative comparisons can express structures at a narrower level of details (finer) than the Must-Link (ML) and Cannot-Link (CL) constraint that are commonly used for semi-supervised clustering. They are also particularly useful in contexts where the granting of a ML or CL constraint is difficult because the granularity of real clustering is unknown. The SKLR algorithm is a break down according to the following steps in Algorithm 3.

Algorithm 2 :SKMS

```
Input:
$D$ : a data set: D
$(Con_{=})$: a set of must-link constraint
$(Con_{\neq})$:a set of cannot-link constraint
$\gamma$: constant distance factor

1: Map the data to a kernel space (kernel).
2: Apply the mean shift on non Approved data.
3: Select the $\sigma$ parameter for the initial kernel Gaussian
   function with minimizing the divergence metric log-det.
4: Calculate the initial matrix
5: Calculate the matrix $K$of kernel $(n*n)$ of low-rank $K_0$
6: Select the mean shift bandwidth parameter k by using $K$
   matrix and constraint$(Con_{=})$
7: Repeat : for $X_i \in D, i = 1...n$
8: Application of the SKMS algorithm to assign data to clusters.
9: until all the constraint are satisfied
   Return : cluster labels
```

Algorithm 3 : SKLR

```
Input:
initial $(n*n)$ kernel matrix $K_0$
$C_{neq}$ et $C_{eq}$: set of relative comparisons
$\gamma$: constant distance factor

1: Find low-rank representation
2: Calculate $(n*n)$ the $K$ kernel matrix of low-rank $K_0$
3: using incomplete Cholesky decompositio: Find $(n*r)$
   column orthogonal matrix $Q$
4: Apply the transformation $\hat{M} \leftarrow Q^{\top}MQ$ S on all matrices
5: Initialize the kernel matrix $\hat{K} \leftarrow \hat{K}_0$
6: Repeat
7: (1) Select an unsatisfied constraint $C \in C_{neq} \bigcap C_{eq}$

8: (2) Apply Bregman projection
9: Until all the constraint are satisfied
   Return $K \leftarrow Q\ K \succ Q$
```

V. Experiment and results discussion

The performance is determined using Rand Index (RI) peer counting methods to measure the influence of stress integration in these algorithms to guide and improve their performance. The comparative study is conducted on 6 selected medical data sets from the UCI repository [13], the characteristics of which are summarized in the TableI.

datasets	#instances	#variables	# class
Pima	768	8	2
Bupa	345	6	2
pancreatic	181	6771	2
heartstatlog	270	13	2
New-thyroid	215	15	3
dermatologie	358	34	6

TABLE I
BENCHMARK DATA SETS USED IN THE EXPERIMENTS

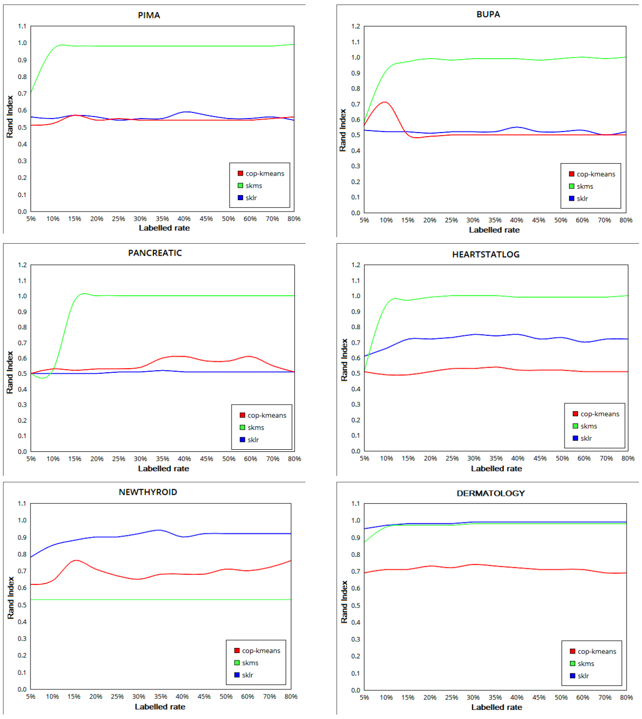

Fig. 3. Clustering results with different constraint

The figure Fig.3, represents the performance graph of the algorithms tested on the selected benchmark. In *pima* dataset, the best performance is recorded by the SKMS algorithm with 15% of labeled data. Observing the *bupa* dataset, we note that the SKLR algorithm records a perfect performance equals to 1 with a small amount of labeled data equals to 15%. On the other hand, the SKLR algorithm remains almost stable with an average value that does not exceed 0.53. The performance of the cop-kmeans algorithm has reached a maximum of 0.71, with some stability of performance for each dataset. In Fig.3, we continue to note the best performance of the SKMS algorithm, the groupings are identical the agreement rate is exactly 1, from 20% of labeled data. In parallel, the SKLR and cop-kmeans algorithms maintain their average performance throughout the experiment. For the *heartstatlog* dataset, we note that the performance of the SKMS algorithm always keeps the best results, while a decrease in SKLR performance is recorded. The red curve that represents the performance of cop-kmeans remains low for the different constraint rates.

For the *dermatology* dataset, we notice that the SKLR curve is identical with the SKMS curve, with high performances at the beginning (10% to15%). Compared to the SKMS method, the cop-kmeans algorithm remains stable with a value not exceeding 0.74 as the best performance recorded. Here, we can say that the SKLR algorithm is the best for this multi-class dataset.

The performance of the algorithms changes completely when using the *new-thyroid* dataset. Indeed, compared to previous experiments, we have found from the graph (Fig.3), that the SKMS results take a constant and average value; which indicates that the algorithm is not affected by the added constraint. On the other hand, the SKLR algorithm keeps the best results. For cop-kmeans algorithm, their best performance is obtained with a high rate of labeled data. We deduce from this analysis that the best algorithm is SKLR for the *new-thyroid* dataset.

We can say according to the experiments carried out, that the algorithms studied achieve good clustering results. We can also mention through these experiments that the quantity and quality of constraint that are created in a random way allow to improve the performance of these methods and they have a direct impact on the results. One can say that at the end of this study, the results of the comparisons revealed that the best compromise is achieved by the SKMS algorithm compared to SKLR and cop-kmeans.

VI. CONCLUSION

In light of the realized comparative study, we can conclude from the obtained results that the use of these 3 methods is very promising, where, we clearly observe by increasing our requirements on similarity, we still manage to extract information to enrich, and get better performance.

The different tracks explored during this work led us to consider many perspectives. We present here those which seem to us the most promising: One of the short-term perspectives is to carry out a comparative study of the different types of constraint and the best way of applying them in order to improve the efficiency of the methods. In addition, in a long-term perspective, the interest will be focused on the automatic segmentation and annotation of structures in biomedical images [26], which are essential tasks for a multitude of key applications including assisted diagnosis, pathology monitoring and clinical research. The process of segmentation is very complex, due in particular to the low contrast, the superposition of regions of interest and noise, typically present in medical images. By the application of the semi-supervised clustering with constraint techniques, it will be possible to automatically label the regions of interest in the image or volume to be segmented. The contribution of the constraint will guide the algorithm for an efficient and targeted segmentation.

References

[1] Ian Davidson and SS Ravi, "Clustering with constraints: Feasibility issues and the k-means algorithm," in *Proceedings of the 2005 SIAM international conference on data mining*. SIAM, 2005, pp. 138–149.

[2] Mikhail Bilenko, Sugato Basu, and Raymond J Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the twenty-first international conference on Machine learning*. ACM, 2004, p. 11.

[3] Sugato Basu, Ian Davidson, and Kiri Wagstaff, *Constrained clustering: Advances in algorithms, theory, and applications*, CRC Press, 2008, pp. 4-11.

[4] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall, "Learning a mahalanobis metric from equivalence constraints," *Journal of Machine Learning Research*, vol. 6, no. Jun, pp. 937–965, 2005.

[5] Kiri Wagstaff, Claire Cardie, Seth Rogers, Stefan Schrödl, et al., "Constrained k-means clustering with background knowledge," in *ICML*, 2001, vol. 1, pp. 577–584.

[6] Zhengdong Lu and Todd K Leen, "Semi-supervised learning with penalized probabilistic clustering," in *Advances in neural information processing systems*, 2005, pp. 849–856.

[7] Kais Allab and Khalid Benabdeslem, "Sélection de contraintes pour la classification topologique semi-supervisée," in *Conférence Francophone d'Apprentissage CAP'11*, 2011, pp. 39–54.

[8] Haitao Gan, Nong Sang, Rui Huang, Xiaojun Tong, and Zhiping Dan, "Using clustering analysis to improve semi-supervised classification," *Neurocomput.*, vol. 101, pp. 290–298, Feb. 2013.

[9] Tomoya Sakai, Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama, "Semi-supervised classification based on classification from positive and unlabeled data," in *Proceedings of the 34th International Conference on Machine Learning*, Doina Precup and Yee Whye Teh, Eds., International Convention Centre, Sydney, Australia, 06–11 Aug 2017, vol. 70 of *Proceedings of Machine Learning Research*, pp. 2998–3006, PMLR.

[10] Marek Śmieja and Magdalena Wiercioch, "Constrained clustering with a complex cluster structure," *Advances in Data Analysis and Classification*, vol. 11, no. 3, pp. 493–518, Sep 2017.

[11] Saket Anand, Sushil Mittal, Oncel Tuzel, and Peter Meer, "Semi-supervised kernel mean shift clustering," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 6, pp. 1201–1215, 2014.

[12] Ehsan Amid, Aristides Gionis, and Antti Ukkonen, "A kernel-learning approach to semi-supervised clustering with relative distance comparisons," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2015, pp. 219–234.

[13] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, "Uci repository of machine learning databases," 1998.

[14] Sugato Basu, Ian Davidson, and Kiri Wagstaff, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, Chapman & Hall/CRC, 1 edition, p 4-11, 2008.

[15] Sophia Roberts, G Gisler, and JAMES Theiler, "Spatio-spectral image analysis using classical and neural algorithms," *Smart Engineering Systems: Neural Networks, Fuzzy Logic, and Evolutionary Programming*, vol. 6, pp. 425–430, 1996.

[16] ULISES CORTÉS and JAVIER BÉJAR, "Experiments with domain knowledge in unsupervised learning: Using and revising theories," *Computación y Sistemas*, vol. 1, no. 003, 1969.

[17] Sugato Basu, Arindam Banerjee, and Raymond Mooney, "Semi-supervised clustering by seeding," in *In Proceedings of 19th International Conference on Machine Learning (ICML-2002*. Citeseer, 2002.

[18] Nizar Grira, Michel Crucianu, and Nozha Boujemaa, "Unsupervised and semi-supervised clustering: a brief survey," 09 2005.

[19] Sugato Basu, Arindam Banerjee, and Raymond J Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proceedings of the 2004 SIAM international conference on data mining*. SIAM, 2004, pp. 333–344.

[20] Sugato Basu, Mikhail Bilenko, Arindam Banerjee, and Raymond J Mooney, "Probabilistic semi-supervised clustering with constraints," *Semi-supervised learning*, pp. 71–98, 2006.

[21] Tetsuya Yoshida, "A graph-based approach for semisupervised clustering," *Comput. Intell.*, vol. 30, no. 2, pp. 263–284, May 2014.

[22] David Chatel, *Semi-supervised clustering in graphs*, Ph.D. thesis, LILLE University - Centre de Recherche en Informatique, Signal et Automatique de Lille (CRIStAL - UMR CNRS 9189), 2017.

[23] Fan RK Chung, *Spectral graph theory*, Number 92. American Mathematical Soc., 1997.

[24] James MacQueen et al., "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Oakland, CA, USA., 1967, vol. 1, pp. 281–297.

[25] Yizong Cheng, "Mean shift, mode seeking, and clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 8, pp. 790–799, Aug. 1995.

[26] Nofech-Mozes S Martel AL. Peikari M, Salama S, "A cluster-then-label semi-supervised learning approach for pathology image classification," *Scientific Reports*, 2018.