Active seed selection for constrained clustering

Viet-Vu Vu, Nicolas Labroche

ACTIVE SEED SELECTION FOR CONSTRAINED CLUSTERING

VIET-VU VU AND NICOLAS LABROCHE

Abstract. Active learning for semi-supervised clustering allows algorithms to solicit a domain expert to provide side information as instances constraints, for example a set of labeled instances called seeds. The problem consists in selecting the queries to the expert that are likely to improve either the relevance or the quality of the proposed clustering. However, these active methods suffer from several limitations: (i) they are generally tailored for only one specific clustering paradigm or cluster shape and size, (ii) they may be counter-productive if the seeds are not selected in an appropriate manner and, (iii) they have to work efficiently with minimal expert supervision. In this paper, we propose a new active seed selection algorithm that relies on a k-nearest neighbors structure to locate dense potential clusters and efficiently query and propagate expert information. Our approach makes no hypothesis on the underlying data distribution and can be paired with any clustering algorithm. Comparative experiments conducted on real data sets show the efficiency of this new approach compared to existing ones.

1. Introduction

Semi-supervised clustering algorithms have gained a lot of attention from the clustering community, as they promise to improve the relevance and the efficiency of traditional methods [21] thanks to the introduction of an expert domain knowledge. This side information can be either provided as class labels for a small set of instances, also called *seeds* [6, 5], or as pairwise constraints between instances [36, 48]. When paired with semi-supervised clustering algorithms, active learning provides an efficient way to solicit an expert to provide the value of a class label or a relation between instances based on what should be the most appropriate for the clustering. However, some studies [46, 47] have shown that constraints or seeds that are not properly chosen can be counterproductive and lead to poor clustering performance, even in the case when the answer of the expert is good.

In this paper, we are interested more specifically in active seed based semi-supervised clustering algorithms that solicit the expert to retrieve class labels. For the time being, research conducted in the field have mainly focused

Key words and phrases. active learning, seed selection, seed based clustering, k-nearest neighbor graph.

on adapting well-known clustering methods to this new semi-supervised context, with the objective to either guide the exploration of the search space to relevant solutions, or to overcome some inherent limitations of clustering algorithms. However, these methods do not address the problem of how selecting the most appropriate seeds while minimizing the expert solicitations. While numerous researches have been conducted in the context of active semi-supervised classification [38], only few methods have been proposed in the clustering context. Moreover, the existing methods are limited by hypothesis on the underlying data distribution and on the shape and sizes of expected clusters [42] or tailored for specific algorithms [51].

To this aim, this paper describes a new active seed selection algorithm, that can be paired with any seed-based clustering algorithm. The idea is to handle the diversity of shapes of clusters with a k-nearest neighbors graph structure to identify the regions of data space in which requesting the expert for labeled instances. Moreover, this graph allows us to query the expert in dense regions of the data sets where the labels provided by the expert can be easily propagated to the neighbors. Experiments on real data sets show the efficiency of our approach with either prototype or density based seed clustering algorithms, its ability to reduce the number of expert queries and finally its robustness to the parameter k in the k-nearest neighbors graph.

This paper is organized as follows: Section 2 presents an overview of the semi-supervised literature and presents the main active seed-selection methods. Then, Section 3 describes our new active seed selection method based on a k-nearest neighbors graph. Section 4 describes the experiments that were carried out and discusses the results. Finally, Section 5 presents the conclusions and perspectives of this research work.

2. Related works

This section first recalls some of main algorithms in the literature of semi-supervised clustering and stresses the sensitivity these approaches to the quantity and the quality of the side information that is provided. Second, this section describes some of the main active seed selection approaches that can help an expert to efficiently - e.g. with minimum expert solicitation - feed semi-supervised clustering algorithm with domain knowledge.

Semi-supervised clustering algorithms aims at extracting patterns from data sets, based on a partial supervision provided by a domain expert [5]. This kind of methods have been shown to be of great interest in the clustering community as recalled in [21]. Indeed, these methods bridge the gap - to a certain extent - between clustering that is a purely unsupervised process, and classification that is a completely supervised learning process. Semi-supervised clustering [49] in this respect allows to add external expert knowledge that may be available to a clustering algorithm, generally as

instance level constraints. There are two main types of instance level constraints: either class labels provided for some instances (also called seeds) [5, 6] or pairwise constraints between instances [36, 48]. There are two types of pairwise constraints: Must-Link (ML) constraints indicate that 2 points should belong to the same cluster while Cannot Link (CL) constraints indicate that 2 points should not be in the same cluster. Other constraints can be found in the literature, such as group-level constraints that are used to overcome constraints consistency problem and allow to express multiple multiple pairwise constraints from a single group constraint [30, 15, 31, 29], or as constraints on the compacity and separability of clusters [13] or such as order preferences on distance [28]. Some works also propose to express preferences on the order of the attributes to guide the search for an efficient metric toward a preferred subset of possible Mahalanobis distances [40, 50].

Most of the research effort until now in the semi-supervised field has been conducted to adapt well-known clustering methods to this new semi-supervised context, with the objective to either guide the exploration of the search space to solutions that are more relevant to the user, or to overcome some inherent limitations of clustering algorithms. For example, seed k-means (SKM) [6, 12] or seed fuzzy c-means [7, 32, 33] allow to reduce the sensitivity of these methods to their initial partition. Similarly, seeds have been used to estimate distinct local density parameters in density-based algorithms like SSDBSCAN, HISSCLU [26, 10].

Additionally, these semi-supervised approaches can be divided in three main categories depending on how they use constraints. Strict enforcement methods explicitly ensure that ML and CL constraints are not violated during the clustering process [6, 49], but as a consequence they sometimes fail to produce a clustering. Soft enforcement methods add a penalization term in the objective function to favor the solutions that best fit constraints [33, 11, 3]. Finally, some other method performs a soft enforcement that relies on the learning of an adapted metric space that minimizes the number of violated constraints [54, 8, 24, 4, 28].

However, all these methods do not address the problem of how selecting the most appropriate constraints for their needs. While numerous researches have been conducted in the context of active query selection for classification [38], very few methods have been proposed in the clustering context. First attempts have been conducted to evaluate the quality of constraints or their utility [14, 45]. Then, most of the existing methods are limited by hypothesis on the underlying data distribution and on the shape and sizes of expected clusters [42].

As stated in [41], and contrary to traditional classification, active learning is *"a selective sampling technique where the learning protocol is in control of*

the data to be used for training". Although active learning for classification has been widely studied [37, 38], it is still under investigation in the clustering community since the advent of semi-supervised clustering [42] and a need for expert interactivity in real-world applications. Two main families of methods can be identified [41]: either *Query-by-Committee* approaches that rely on several classifiers to decide which query to ask the expert for or *Uncertainty Sampling* approaches that rely on a single classifier and select the example for a potential user query based on those of the examples that exhibit the smallest confidence from the classifier. In the case of clustering, most of the active learning approaches rely on the uncertainty sampling as they decide on the query to ask the expert based on the assignment of points on which an error is more likely to penalize the objective function. As an example, spectral clustering methods have been particularly studied in recent years [9, 51, 52, 53]. In this case, the general principle consists in querying the constraints which maximally reduce the expected error of the clustering algorithm. However, these active learning methods are generally tailored to work specifically with only one single clustering method in mind. Our objective is to propose general methods that can fit any clustering context.

The problem of selecting the best seeds in the context of clustering algorithms has already been partially covered by papers related to the problem of initialization of centers in k-means like algorithms [1, 7, 17, 20]. As recalled by [34], this problem has been deeply studied but one can identify four major approaches to initialize the centers in k-means like approaches: the random creation of the initial partition, the classical Forgy method as reported by [2] in which initial seeds are randomly selected (and then all points are assigned to the nearest seed), the MacQueen method in which, similarly to Forgy, seeds are chosen randomly, but then each time a point is assigned to a seed, the corresponding cluster center is updated, and finally the Kaufman approach [23] in which the first seed is the center of the data set and all the other seeds are selected according to a criterion that depends on the number of data in the neighborhood of the seed candidate and the distance to the seeds that are already selected.

Other approaches like [19, 39] also select the seeds based on their distance to the set of seeds already selected. More precisely, in [39], authors propose two heuristics that either maximize the sum of the distance or the minimal distance to the existing seeds. Finally, [18] initializes the first seed as the center of the data set and then selects randomly other point that are averaged with the center coordinate with an appropriate weight to cover the entire data set while being more resistant to outliers.

All the previous methods (random selection or maximization of the distance to already selected seeds) allow an efficient coverage of the data space and some of the approaches like [23] also take into account a density measure (number of data points in the neighborhood) to choose from all the possible distant seeds. More recently, [27] propose an active learning method based

on Minimum Spanning Tree (MST) to deal with multi-densities and imbalanced data sets while minimizing the number of experts solicitations. The main idea is to build a MST from a sample of the data set and then cut this graph into k clusters. Then a query is asked to the expert for each of the k parts of the graph and the label is propagated to all the neighbors based on thresholds learned from each cluster. This method is very costly in terms of complexity with $O(n^2 logn)$ computations for a data set with n object, hence the proposal for a sampling. Our approach is based also on a propagation mechanism, but contrary to the previous method, complexity is limited in the worst case to $O(n^2)$ and has fewer parameter to set. Indeed, it can adapt automatically to different densities without any threshold parameters.

We now describe with more detail the two main methods named S-Min-Max and S-Min-Max-D that have been proposed to allow the active selection of seeds by an expert in the context of semi-supervised clustering [42].
Min-Max approach. the objective of the *Min-Max* approach is to build a set of seeds Y from a data set X such that seeds in Y are evenly spaced and produce a good coverage of the data space [42]. Moreover, the method aims at minimizing the annotation effort of the expert and thus tries to minimize the number of seeds that represent the same cluster. Initially, as there is a priori no information about the data set, the first seed of the set Y is chosen randomly among data points in X. Then, the next seed y_{new} has to maximize its minimal distance to the set of seeds already selected as shown in the following Equation 1.

$$y_{new} = argmax_{x \in X-Y} \left(min_{y \in Y} D(x, y) \right) \tag{1}$$

where y_{new} denotes the new point to be added to the seeds set Y and where D) denotes a metric defined in the data space of points X (for example D could be an euclidian distance or a Mahanalobis distance if we compare R^m vectors, a Levenshtein distance if we compare sequences ...).

The active seed selection algorithm based on the Min-Max approach, called S-Min-Max, is an iterative process where, at each step, a new seed candidate y_{new} (as determined by Equation 1) is proposed to the expert to be labeled. As in any active learning system, the expert is supposed to be able to answer to all the queries of the system. The iterative process stops when the experts decides to or when all points in X have been explored. However, because of its formulation, S-Min-Max is sensitive to outliers that naturally maximize their minimal distance to any cluster. Moreover, when clusters are elongated, choosing the farthest point from a seed can lead to the selection of a point in the same cluster. In other words, Min-Max method does not guaranty that seeds are located around cluster centers, which can be problematic with k-means like clustering algorithms. Finally, the results depend heavily on the first selected seed as well as the shape and the size of

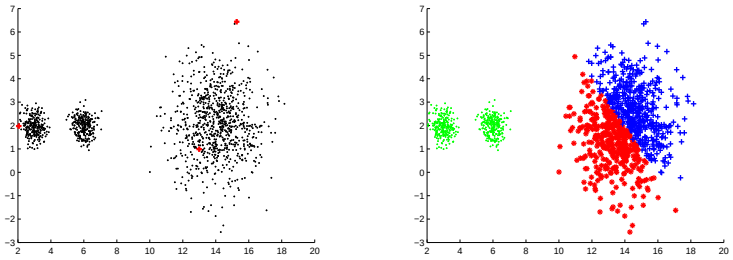

FIGURE 1. Limit of the *Min-Max* approach. The seeds selection process heavily depends on the first seed that is randomly selected and the size and the shape of clusters. **Left**: data set with 3 clusters and selection of 3 seeds with the *Min-Max* approach, two being in the largest cluster. **Right**: resulting clustering with the seed k-means algorithm.

the clusters and the number of seeds available for each cluster. As an example, Figure 1 illustrates a case where a seed is provided in the surrounding of a cluster and not near its center, which causes an erroneous convergence of seed k-means algorithm in this case.

Min-Max approach based on local density. the previous *S-Min-Max* approach relies on the hypothesis that clusters are mainly hyperspherical, so that maximizing the distance between each proposed seed, maximizes the chances that the seeds belong to distinct clusters. However, the method becomes sensitive to the relative position of the first selected seed compared to the expected clusters. The improved *Min-Max Local Density Score* (S-Min-Max-D for short) improves the S-Min-Max approach by initializing the seeds near the clusters centers, that is to say, in dense regions of the data set. In order to determine the potentially interesting regions of the data space with high density, the S-Min-Max-D relies on a k-nearest neighbors graph structure, similarly to what have been proposed for active constraint selection algorithms [43, 44, 45].

The k-nearest neighbors graph. (k-NNG) is a weighted undirected graph, in which each vertex represents a data point, and has at most k edges to its k-nearest neighbors. An edge is created between a pair of vertices, u and v, if and only if the points associated to vertices u and v have each other in their k-nearest neighbors set. The weight $\omega(u, v)$ of the edge between the vertices u and v is defined as the number of common nearest neighbors the two points associated to u and v share, as shown in equation 2 [22]:

$$\omega(u, v) =\mid NN(u) \cap NN(v) \mid \tag{2}$$

where *NN(.)* denotes the set of k-nearest neighbors of the associated point. This similarity measure is interesting since it can adjust automatically the density to different contexts and clusters. Thus, it is possible to extract a local density indicator from a k-NNG with the Local Density Score [25], which is defined as the average, for each point, of the proximity ω with all its neighbors as recalled in Equation 3.

$$LDS(u) = \frac{\sum_{q \in NN(u)} \omega(u, q)}{k} \tag{3}$$

The LDS value of a point x is set in $[0, k-1]$ where k is the number of nearest neighbors. It is defined so that a high value of $LDS(x)$ indicates a high proximity between the point x and its neighbors, i.e. x belongs to dense region of the data space. Similarly, a small value of $LDS(x)$ indicates that x belongs to a transition region between clusters or x is an outlier with far nearest neighbors.

As the proposed S-Min-Max-D aims at focusing on candidate seeds near the centers of potential clusters, the previous S-Min-Max approach is modified by integrating a preliminary filtering of the candidate seeds $Candidate\_Set$ on the basis of their LDS scores and a minimum density threshold ϵ as shown in Equation 4.

$$Candidate\_Set = \{p \in X : LDS(p) \geq \epsilon\} \tag{4}$$

where ϵ is a parameter that has to be set experimentally.

Figure 2 illustrates a $Candidate\_Set$ obtained on the previous data set with 3 clusters (see Figure 1). As expected, seed candidates are selected in dense regions of the data space, near the cluster centers and can thus favor k-means like clustering algorithms.

In this paper, we are interested more specifically in active seed based semi-supervised clustering algorithms that query the expert to retrieve interesting class labels. To this aim, Section **??** describes our new density-based approach that can deal with any cluster shape and can adapt equally to different data densities.

3. Proposed approach for seed selection

Our active learning seed selection method falls into the category of uncertainty sampling approaches, but contrary to these methods that try to find regions of the data space where clustering algorithms are supposed to make assignment mistakes, we focus on determining regions where we are almost certain that the points belong to a single cluster.

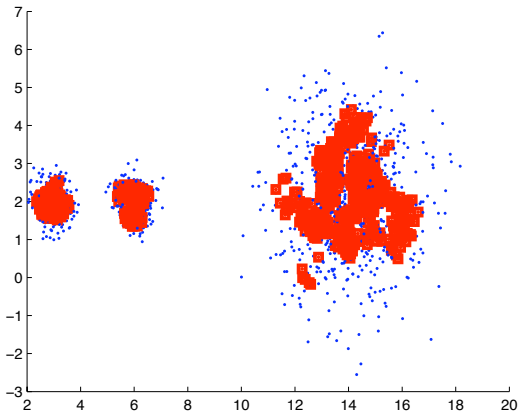

FIGURE 2. Illustration of seed candidates obtained with the S-Min-Max-D algorithm with parameters $k = 50$ and $\epsilon = 25$.

Moreover, our approach avoids the Min-Max scheme that allows for a good coverage of the data space for two main reasons. First, Min-Max methods make the assumption that clusters are hyper-spherical. Even if this works well in the case of k-means-like semi-supervised algorithms, this may as well lead to ask several queries to the expert in case of elongated clusters that are to be detected with density-based semi-supervised clustering algorithms. Second, Min-Max approach is very sensitive to outliers as it chooses as seed candidates, the points that maximize their minimal distance to already found clusters. Again, this may be counter-productive to ask the expert to label outliers.

The solution that we propose relies on the efficient detection of clusters as dense regions of the data space. It is then possible to discover either hyper-spherical or elongated clusters as well. But then, the difficulty, as illustrated with previous semi-supervised density-based clustering algorithms [26], is that the algorithm has to seamlessly adapt to distinct densities to be usable in our context. To this aim, our method relies on a k-nearest neighbors graph structure (k-NNG) as in the previous S-Min-Max-D method [42] to detect the potentially interesting dense regions of the data space. Contrary to this previous method that only uses the k-NNG to filter seed candidates, our new approach takes full advantage of the structure to reduce expert solicitations. Indeed, the interesting regions are represented as connected components of the k-NNG, that is to say, set of points that are highly related with their direct neighbors. In a k-NNG, the weight ω (see Equation 2) reflects this local proximity between two neighbors. In this case, it is important to notice that the expression of the density as a number of common nearest neighbors

rather than a number of points in a given distance related neighborhood allows to be independent of the actual observed distance between points. As a consequence, the k-NNG structure naturally fits our objective of adapting to clusters of distinct densities. Moreover, the k-NNG allows us to query the expert in dense regions of the data sets where the labels provided by the expert can be easily propagated to the neighbors thanks to the connectivity of the graph structure, hence reducing in turn the number of queries and the expert involvement.

In our proposal, an initial set of seeds candidates is determined by considering only the edges of the k-NNG whose weight ω is over a fixed threshold value θ as reflected by the following Equation 5:

$$Candidate\_Set = \{\forall u, v \in X : \omega(u, v) \geq \theta\} \tag{5}$$

where $\theta \in [0, k]$ is a parameter of our method. The lower θ, the more seeds candidates and the sparser regions can be taken into account. Thus it is important to set θ to a compromise, in order not to consider too many candidate seeds. In our experiments, we set θ so that the set of initial seeds covers approximately 70% of the whole data set.

Then, the set of seeds candidates is used to support the active generation of queries to the expert. In order to minimize the annotation effort, each query to the expert should ideally maximize the number of class labels that can be inferred. To that aim, our algorithm builds the connected components of the previous candidate set on the basis of the k-NNG and order the resulting components in the descending order of their cardinality: the larger connected component is ranked first.

As in other active seeds selection methods, the seed selection is an iterative process that stops when the experts decide to, or when the candidate set is empty. It is also assumed that the expert can answer every queries (s)he is asked for. At each iteration, the algorithm chooses randomly a seed candidate from the first ranked connected component that has not been queried, generates the query and retrieves the corresponding class label provided by the expert. The novelty here is that it is possible to naturally propagate the class label to all the candidate seeds that belongs to the same connected component, similarly to what is done in classical supervised learning. Then, in order to avoid soliciting the expert several times for the same potential cluster, each time a candidate seed is labeled, its corresponding connected component is removed from the list of potentially interesting connected component so that no other seed candidate from the same region can be submitted for a query to the expert. As a summary, the main steps of our active seeds selection algorithm are presented in the algorithm 1 hereafter.

Algorithm 1 Active seeds collection method based on a k-NN graph

Require: Data set X, density threshold θ
Ensure: Set of seeds Y
$Y = \emptyset$
$C = \{(u, v) \in k\text{-NNG} : \omega(u, v) \geq \theta\}$
Build the set of connected components from C:
$CC = \{C_1, C_2, \ldots, C_m\}$
repeat
  Randomly select $u \in C_v$ such that $|C_v| = max_{c \in CC}|c|$
  Query the expert to get the class label of u
  if answer exists **then**
    Propagate the class label of u to all points in C_v
    $Y = Y \cup C_v$
    $CC = CC - C_v$
  end if
until $((User\_stop = true)$ **or** $(CC = \emptyset))$
return Y

The complexity of this approach depends heavily on the complexity to build the k nearest-neighbors graph which in turn depends on the dimensionality of the data space. As noted in [43], the complexity can range from $O(n * k)$ with low dimensional data, to $O\left(n * log(n)\right)$ when it is larger but still less than 20 dimensions and to $O(n^2)$ above.

4. Experiments and results

Several experiments have been conducted to assess the quality of our new approach. As our main objective is to propose an active method that makes no hypothesis on the shape or the size of the expected clusters, comparative experiments are performed with either *seed k-means* (SKM) [6] or *SSDBSCAN* [26] semi-supervised clustering algorithms. The section is organized in three main parts: (i) first, it details the experimental protocol for each of our tests, (ii) second, it shows the results related to the quality of our approach based on a clustering evaluation and the number of expert solicitations and, (iii) it discusses the robustness to the parameters k and θ that respectively determines the connectivity of the k-NNG structure and controls the building of the initial set of candidate seeds.

4.1. **Experimental protocol.**

Data sets description. As shown in Table 1, we use 6 well-known real data sets from the Machine Learning Repository [16] named: Iris, Soybean, Protein, Zoo, Thyroid, and LetterIJL and 6 data sets extracted from CALTECH-101 image data set [16], to evaluate our algorithm. These data sets have been chosen because they facilitate the reproducibility of the experiments

TABLE 1. Main characteristics of the real data sets

Data	**#Objects**	**#Attributes**	**#Clusters**
Iris	150	4	3
Soybean	47	34	4
Protein	116	20	6
Zoo	101	16	7
Thyroid	215	5	3
LetterIJL	227	16	3
Test1	240	128	5
Test2	308	128	6
Test3	336	128	7
Test4	428	128	8
Test5	458	128	9
Test6	672	128	10

and because some of them have already been used in constraint-based clustering articles. As can be seen in Table 1, these data sets cover several difficulties for a clustering algorithm. Data set sizes range from very small (47 instances in Soybean) to small (672 instances in Test6). Here, we have preferred to favor small data set because of the complexity of the k-NNG structure that grows with the number of attributes that ranges from 4 (Iris) to 128 (Caltech image data sets). Finally, various cluster numbers and sizes are also investigated.

Compared methods. In the experiments, we denote the methods as follows:

- *S-Random* refers to a complete random seeds selection from the whole data set. It is a naive approach that we use as a baseline for reference purpose ;
- *S-Min-Max* refers to the simple Min-Max approach as formalized by Equation 1 ;
- *S-Min-Max-D* refers the extended Min-Max approach with a preliminary filtering of candidate seeds with a k-NNG ;
- *S-k-NNG* refers to our new approach that identify seed candidates and propagate class labels with a k-NNG.

Evaluation of the results. As all our benchmark data sets contain a class label, we use the Rand Index (*RI* hereafter) [35] in our experiments to evaluate the agreement between the theoretical partition of each data set and the output partition of the evaluated algorithms. *RI* takes values between 0

and 1; $RI = 1$ when the result is the same as the ground-truth. The larger the RI, the better the result.

We also evaluate, for each data set, the number of expert queries needed to collect the set of seeds. For each method, the process of collecting seeds stops when at least one seed was chosen in each of the clusters.

4.2. **Rand Index results.** Figures 3 and 4 present the Rand Index (RI) scores for the four previous active seeds selection methods.

First, it can be noticed that the S-Random method, even if simple, allows for comparable RI scores with the other heuristics on some of our data sets. The S-Random performances decrease mostly on the LetterIJL data set (with SSDBSCAN) or the Caltech image data sets Test1 (with SKM or SSDBSCAN) or Test4 (with SSDBSCAN). This can be explained by the fact that as the expert always provides a good answer, it is sufficient on some of our data sets to improve the clustering quality. However, this relatively good results come at the price of a higher number of expert queries (see Figure 5).

Conversely, when paired with Seed k-means, our approach is always either similar or slightly better in terms of RI than than the other methods. It is interesting to notice that in this case, our approach still performs better than S-Min-Max-D, even if it does not rely on a Min-Max approach that favors k-means like methods. It can be explained by the fact that the k-NNG allows our approach to identify correctly the neighborhood of the cluster centers without a need for a Min-Max heuristic to explore data space.

When paired with SSDBSCAN, our S-k-NNG approach is significantly better than the other methods. Indeed, contrary to the S-Random and S-Min-Max that do not necessarily search for seeds in the dense region of the data set, our S-k-NNG method and S-Min-Max-D benefits from more appropriate seeds in the center of potential clusters to start a density-based clustering algorithm. In this case however, S-Min-Max-D does not allow to label as much instances of each cluster as our new approach that benefits from an efficient propagation mechanism. This mechanism allows for a more accurate labelling of points in each cluster near the decision frontier, which in turn advantages SSDBSCAN that stops the growth of a cluster as soon as an instance with an other cluster label is to be merged in the current cluster. The ability of our method to cover more instances with it propagation mechanism is indirectly illustrated by the number of queries that is lower for the S-k-NNG than the S-Min-Max-D method as shown in Figure 5.

4.3. **Number of question used in active learning process.** As shown in Figue 5, it is interesting to notice that S-Random is generally the less efficient method regarding this particular criterion. Indeed, selecting seeds at random does not allow to find the most appropriate labels for SKM and SSDBSCAN and thus require a lot more "test and try" before converging to a good clustering solution.

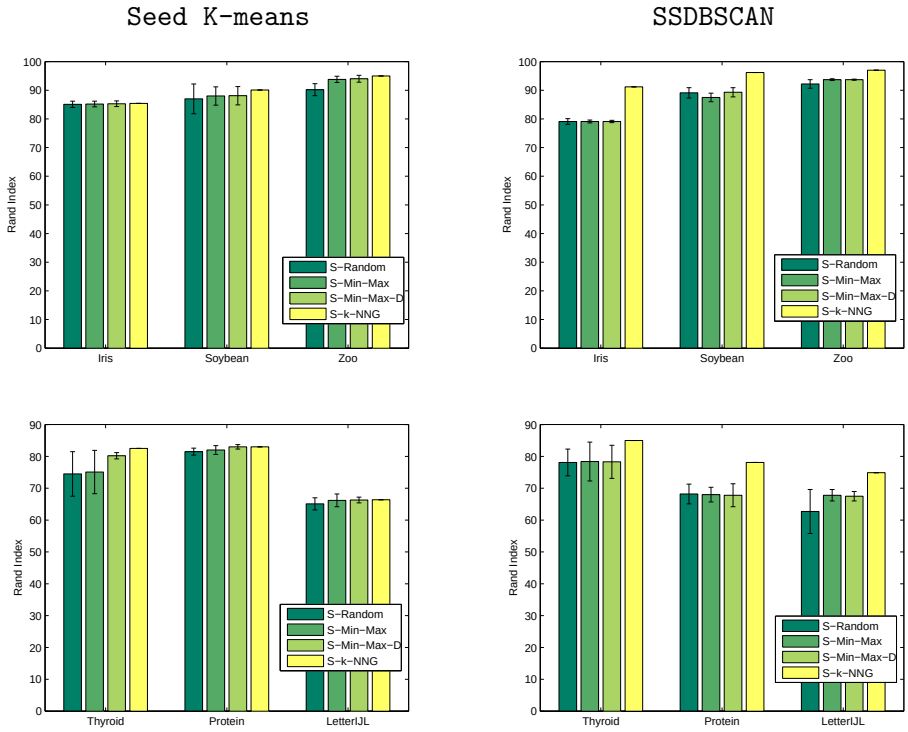

FIGURE 3. Rand Index for 4 methods of seeds selection with Seed K-Means and SSDBSCAN for 6 UCI data sets.

As expected, the S-k-NNG obtains the best results with up to 4 times less questions than the worst approach for the Zoo data set (7 queries versus more than 30 queries for the S-Random approach in this case). This experiment shows the advantage to use a k-NNG graph representation and the effectiveness of our propagation mechanism of class labels according to connected components. It is also interesting to notice that this propagation allows our method to be resistant to unbalanced clustering while other approaches like S-Random and S-Min-Max will ask several queries for the largest clusters. As a conclusion, even if the clustering quality is sometime similar with previous approaches, our new S-k-NNG solicits less the expert on all our benchmark data sets.

4.4. **Robustness and parameter settings.**

Influence of the number of nearest neighbors k. A study has been conducted to evaluate the influence of the number of nearest neighbor k when our S-k-NNG approach is paired with SSDBSCAN. Figure 6 shows the obtained Rand Index for distinct values of k. It can be seen that, for each data set there exists a large interval to find a good value of k. Interestingly, small

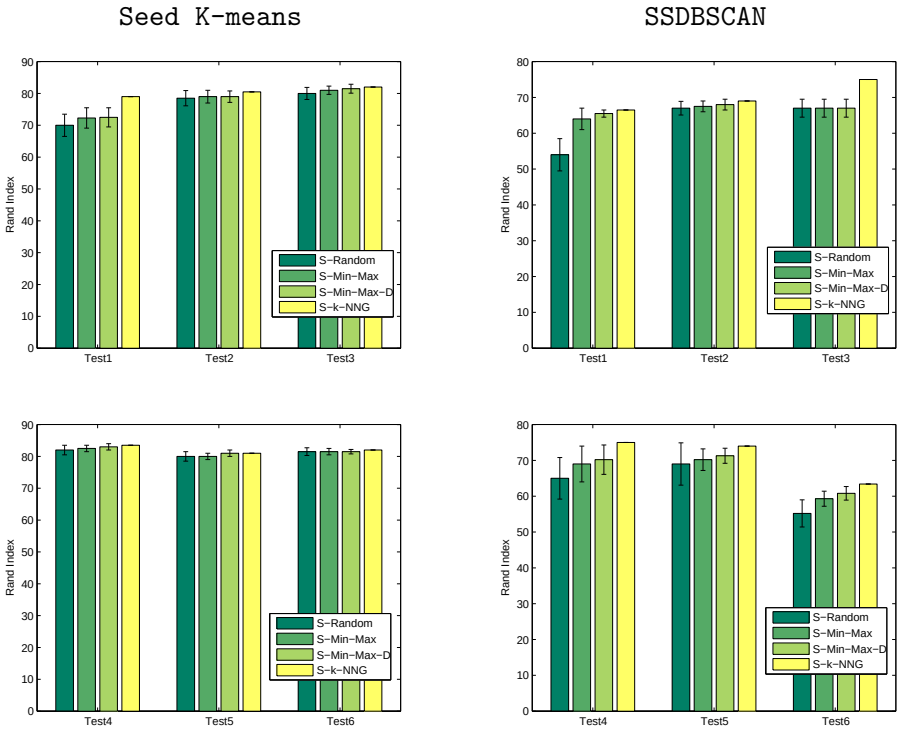

FIGURE 4. Rand Index for 4 methods of seeds selection with Seed K-Means and SSDBSCAN for 6 images data sets.

and large values of k can lead to a decrease in clustering accuracy for some data sets. This can be explained by the fact that a high value aggregate too much distinct clusters while a small value lead to an over-clustering of the data set. This phenomenon is well known in density-based clustering algorithms like DBSCAN where both parameters, the number of neighbors and the size of the neighborhood, are very sensitive and convey the same type of information as our k parameter.

Influence of the parameter θ. The value of θ parameter drives the construction of the initial set of candidate seeds as reported in Equation 5. This parameter is important as it reflects how much of the border of each cluster is filtered before generating the questions. It can be seen as a threshold that allows an expert to avoid asking questions near the decision frontier and should be set experimentally when the experts know that there are probably overlapping clusters. It is set in our experiment such that this set of candidate seeds contains about 70 percents of the whole data set. Our experiments show that it is possible to define, similarly to k, a confidence

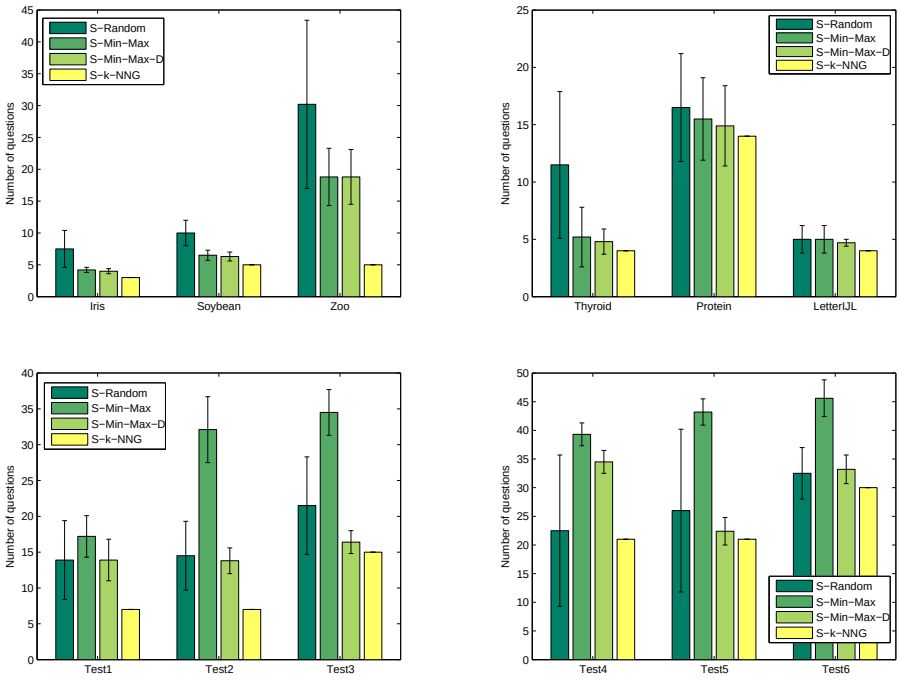

FIGURE 5. Number of question for 4 methods with 12 data sets.

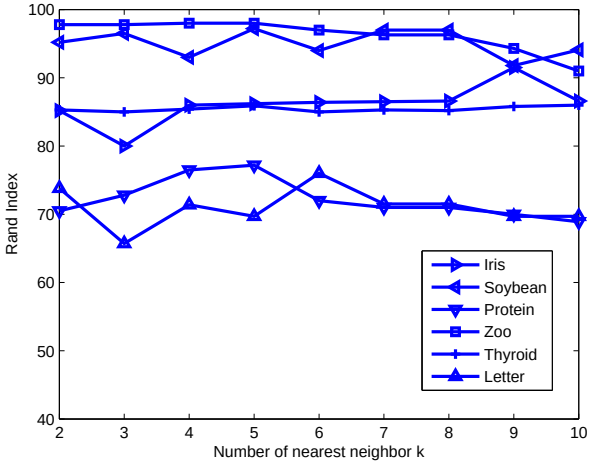

FIGURE 6. Rand Index for 6 UCI data sets with the SSDB-SCAN algorithm paired with our S-k-NNG seeds selection method for $k \in [2, 10]$.

interval where to set the value of θ. Figure 7 plot the Rand Index for several values of $\theta \in [0, k-1]$ and several values of k on our benchmark data sets. Initial values for θ are set in the $[0, k-1]$ interval since this threshold applies on ω values that are also defined in this interval. Contrary to the previous k parameter, results are more difficult to interpret because the value of θ depends on the data set and the value of k. However, it is still possible to observe that our method achieves good results on our benchmark data sets when $\theta \in [\frac{k}{2} - 1, \frac{k}{2} + 2]$.

5. Conclusion and future work

This paper introduces a new active seed selection method named S-k-NNG that is efficient with any kind of semi-supervised clustering algorithms. Similarly to some previous approaches like Min-Max-D, S-k-NNG relies on an underlying k-nearest neighbors graph structure to build the set of initial seed candidates. However, and contrary to previous methods, S-k-NNG also uses the k-nearest neighbors graph to produce a first clustering of the data set based on the density, that allows for an efficient propagation of instance labels provided by an expert. Experiments on real data sets suggest that our new active method is comparable or more efficient than the others when paired with a representative-based clustering like Seed K-Means and that it is much more efficient when paired with a density-based clustering algorithm like SSDBSCAN. Then, our experiments show that our approach allows for less expert solicitations than other methods which is crucial in most real use cases. Finally, our tests show that our approach is robust to its two main parameters, the number of nearest neighbors k and the threshold θ that constraint the construction of the set of initial candidate seeds.

The bottleneck of the S-k-NNG approach, as for the previous Min-Max-D, is the building of the k-nearest neighbors graph which can be costly when the dimensionality of the data space increases. Thus, future work aims at evaluating approximate heuristics to determine efficiently the nearest-neighbors. Other possible extensions of this work concerns an in depth analysis of the behavior of S-k-NNG in an interactive context when all the seeds are not provided before clustering but during the clustering process. This interactivity raises new questions such as how to evolve the parameters k and θ based on previous labels provided by the expert and how to handle label contradictions occurring during time or label propagation when new instances labels are available.

References

[1] M. El Agha and W.M. Ashour. Efficient and fast initialization algorithm for k-means clustering. *I.J. Intelligent Systems and Applications*, 1:21–31, 2012.

[2] M.R. Anderberg. *Cluster Analysis for Applications*. Academic Press, New York, 1973.

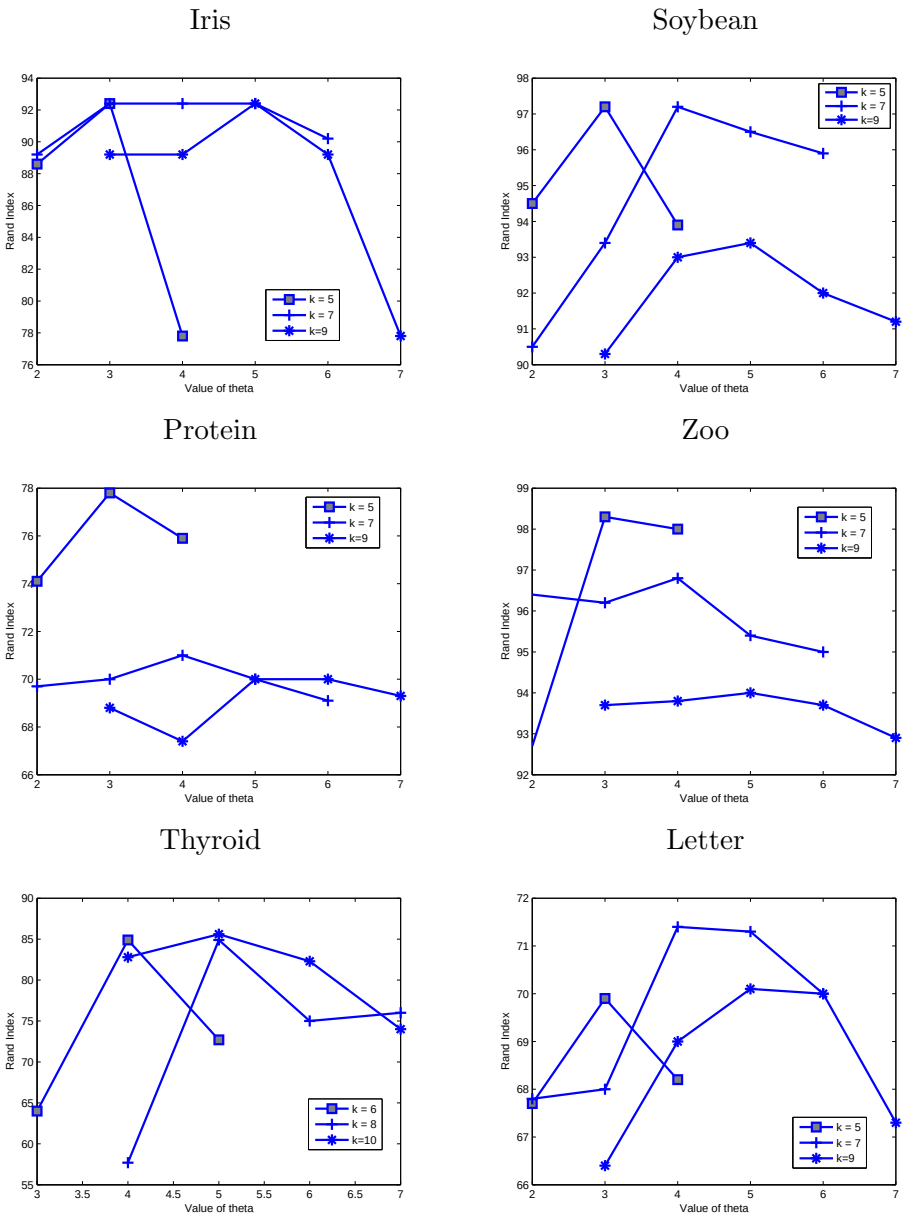

FIGURE 7. Rand Index for several k and θ parameters values on 6 benchmark data sets when the active S-k-NNG method is paired with SSDBSCAN semi-supervised clustering.

[3] V. Antoine and N. Labroche. Classification évidentielle avec contraintes d'étiquettes. In *Proc. of EGC Conference*, pages 125–136, 2015.
[4] Aharon Bar-Hillel, Tomer Hertz, Noam Shental, and Daphna Weinshall. Learning a mahalanobis metric from equivalence constraints. *J. Mach. Learn. Res.*, 6:937–965, December 2005.
[5] S. Basu, I. Davidson, , and K. Wagstaff, editors. *Constrained Clustering: Advances in Algorithms*. Theory, and Applications. Chapman & Hall/CRC, 2008.
[6] Sugato Basu, Arindam Banerjee, and R. Mooney. Semi-supervised clustering by seeding. In *In Proceedings of 19th International Conference on Machine Learning (ICML-2002)*, 2002.
[7] A. M. Bensaid, L. O. Hall, J. C. Bezdek, and L. P. Clarke. Partially supervised clustering for image segmentation. *Pattern Recognition*, 29(5), 1996.
[8] M. Bilenko, , S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *Proc. of the 21st ICML Conference*, page 11. ACM, 2004.
[9] Z. Bodó, Z. Minier, and L. Csatõ. Active learning with clustering. *Journal of Machine Learning Research*, 16:127–139, 2011.
[10] C. Böhm and C. Plant. Hissclu: a hierarchical density-based method for semi-supervised clustering. In *In Proc. of the 11th International Conference on Extending Database Technology (EDBT)*, 2008.
[11] A. Bouchachia and W. Pedrycz. A semi-supervised clustering algorithm for data exploration. In *Proc. Internat. Fuzzy Systems Association World Congress*, pages 328–337, 2003.
[12] Y. Dang, Z. Xuan, L. Rong, and M. Liu. A novel initialization method for semi-supervised clustering. In *Proc. of 4th Int. Conf. KSEM*, 2010.
[13] I. Davidson and S.S. Ravi. Agglomerative hierarchical clustering with constraints: Theoretical and empirical results. In *Proceeding of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 59–70, 2005.
[14] I. Davidson, K.L. Wagstaff, and S. Basu. Measuring constraints-set utility for partitional clustering algorithms. In *Proceeding of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pages 577–584, 2006.
[15] A. Dubey, I. Bhattacharya, and S. Godbole. A cluster-level semi-supervision model for interactive clustering. In *Proceedings of the 2010 European Conference on Machine Learning and Knowledge Discovery in Databases: Part I*, ECML PKDD'10, pages 409–424, Berlin, Heidelberg, 2010. Springer-Verlag.
[16] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *In Proc. Workshop on Generative-Model Based Vision, IEEE CVPR 2004*, 2004.
[17] M.A. Hasan, V. Chaoji, S.Salem, and M.J. Zaki. Robust partitional clustering by outlier and density insensitive seeding. *Pattern Recognition Letters*, 30(11):994–1002, 2009.
[18] J. Heer and E. Chi. Identification of web user traffic composition using multi-modal clustering and information scent. In *Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining*, pages 51–58, 2001.
[19] R.E. Higgs, K.G. Bemis, I.A. Watson, and J.H. Wikel. Experimental designs for selecting molecules from large chemical databases. *J. Chem. Inf. Comput. Sci*, 37:861–870, 1997.
[20] D. S. Hochbaum and D. B. Shmoys. A best possible heuristic for the k-center problem. *Mathematics of operations research*, 10(2):180–184, 1985.
[21] Anil K. Jain. Data clustering; 50 years beyond k-means. *Pattern Recognition Letters*, pages –, 2009.

[22] R. A. Jarvis and E. A. Patrick. Clustering using a similarity measure based on shared near neighbors. *IEEE Transactions on Computer*, 22(11):1025–1034, 1973.

[23] L. Kaufman and P.J. Rousseeuw. Finding groups in data: An introduction to cluster analysis. In *John Wiley and Sons*, 1990.

[24] D. Klein, S. Kamvar, and C. Manning. From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering. In *Proc. of the 19th Int. Conf. on Machine Learning (ICML '02)*, pages 307–314, 2002.

[25] D.-D. Le and S. Satoh. Unsupervised face annotation by mining the web. In *In Proc. of the IEEE International Conference on Data Mining (ICDM)*, pages 383–392, 2008.

[26] L. Lelis and J. Sander. Semi-supervised density-based clustering. In *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*, pages 842–847, Dec 2009.

[27] Mingwei Leng, Jianjun Cheng, Jinjin Wang, Zhengquan Zhang, Hanhai Zhou, and Xiaoyun Chen. Active semisupervised clustering algorithm with label propagation for imbalanced and multidensity datasets. *Mathematical Problems in Engineering*, 2013.

[28] E. Y. Liu, Z. Guo, X. Zhang, V. Jojic, and W. Wang. Metric learning from relative comparisons by minimizing squared residual. In *Proc. IEEE 12th ICDM*, pages 978–983, 2012.

[29] H. Liu and Y. Fu. Clustering with partition level side information. In *Data Mining (ICDM), 2015 IEEE International Conference on*, pages 877–882, Nov 2015.

[30] A. K. Jain M. H. C. Law, A. Topchy. Clustering with soft and group constraints. In *Proc. of Joint IAPR International Workshops, SSPR 2004 and SPR 2004*, pages 662–670, 2004.

[31] B. M. Nogueira, A. M. Jorge, and S. O. Rezende. Hcac: Semi-supervised hierarchical clustering using confidence-based active learning. In *Proc. of 15th Int. Conf. Discovery Science*, pages 139–153, 2012.

[32] W. Pedrycz. Algorithm of fuzzy clustering with partial supervision. *Pattern Recognition Letters*, 3:13–20, 1985.

[33] W. Pedrycz and J. Waletzky. Fuzzy clustering with partial supervision. *IEEE Transactions on systems, Man, and Cybernetics*, 27(5):787–795, 1997.

[34] J.M. Pensa, J.A. Lozano, and P. Larranaga. An empirical comparison of four initialization methods for the k-means algorithm. *Pattern Recognition Letters*, 20:1027–1040, 1999.

[35] W.M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, 1971.

[36] M. Schultz and T. Joachims. Learning a distance metric from relative comparisons. In *Proc. of NIPS Conference*, 2004.

[37] B. Settles. Active learning literature survey. Technical Report Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[38] B. Settles. Active learning literature survey. Technical Report 1648, University of Wisconsin-Madison, 2010.

[39] M. Snarey, N.K. Terrett, P. Willet, and D.J. Wilton. Comparison of algorithms for dissimilarity-based compound selection. *J. Mol. Graphics and Modelling*, 15:372–385, 1997.

[40] J. Sun, W. Zhao, J. Xue, Z. Shen, and Y. Shen. Clustering with feature order preferences. *Intelligent Data Analysis*, 14:479–495, 2010.

[41] K. Tomanek and U. Hahn. Semi-supervised active learning for sequence labeling. In *Proc. of ACL '09 and 4th Int. Joint Conf. on Natural Language Processing*, pages 1039–1047, 2009.

[42] V.-V. Vu, N. Labroche, and B. Bouchon-Meunier. Active learning for semi-supervised k-means clustering. In *In Proc. of the 22nd IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 12–15, 2010.

[43] V.-V. Vu, N. Labroche, and B. Bouchon-Meunier. Boosting clustering by active constraint selection. In IOI press, editor, *In Proc. of the 19th European Conference on Artificial Intelligence (ECAI-2010)*, pages 297–302, 2010.

[44] V.-V. Vu, N. Labroche, and B. Bouchon-Meunier. An efficient active constraint selection algorithm for clustering. In IOI press, editor, *In Proc. of the 20th IEEE International Conference on Pattern Recognition (ICPR-2010)*, pages 2969–2972, 2010.

[45] V.-V. Vu, N. Labroche, and B. Bouchon-Meunier. Improving constrained clustering with active query selection. *Pattern Recognition*, 4(45):1749–1758, 2012.

[46] K. Wagstaff. When is constrained clustering beneficial, and why? In *Proc. of the 21st Nat. Conf. on Artificial Intelligence (AAAI)*, 2006.

[47] K. L. Wagstaff. Value, cost, and sharing: Open issues in constrained clustering. In *In Proc. of the 5th Int. Workshop on Knowledge Discovery in Inductive Databases*, pages 1–10, 2007.

[48] K. L. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proc. of the 17th Int. Conf. on Machine Learning (ICML 00)*, pages 1103–1110, 2000.

[49] K. L. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl. Constrained k-means clustering with background knowledge. In *Proc. of the 18th Int. Conf. on Machine Learning (ICML 01)*, pages 577–584, 2001.

[50] J. Wang, S. Wu, and G. Li. Clustering with instance and attribute level side information. *Int. Journal of Computational Intelligence Systems*, 3(6):770—785, 2010.

[51] X. Wang and I. Davidson. Active spectral clustering. In *Proc. of Int. Conf. on Data Mining (ICDM 2010)*, 2010.

[52] X. Wang and I. Davidson. Flexible constrained spectral clustering. In *Proceeding of the Conference on Knowledge Discovery and Data Mining*, pages 563–572, 2010.

[53] Fabian L. Wauthier, Nebojsa Jojic, and Michael I. Jordan. Active spectral clustering via iterative uncertainty reduction. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 1339–1347, New York, NY, USA, 2012. ACM.

[54] E. P. Xing, M. I. Jordan, S. Russell, and A. Y. Ng. Distance metric learning with application to clustering with side-information. In *Proc. of NIPS Conference*, pages 505–512, 2002.