Journal of Computer Science and Cybernetics, V.35, N.4 (2019), 373–384
DOI 10.15625/1813-9663/35/4/14123

CHOOSING SEEDS FOR SEMI-SUPERVISED GRAPH BASED CLUSTERING

CUONG LE[1], VIET-VU VU[1,*], LE THI KIEU OANH[2], NGUYEN THI HAI YEN[3]

[1] *VNU Information Technology Institute, Vietnam National University, Hanoi*
[2] *University of Economic and Technical Industries*
[3] *Hanoi Procuratorate University*
[*] *vuvietvu@vnu.edu.vn*

Abstract. Though clustering algorithms have long history, nowadays clustering topic still attracts a lot of attention because of the need of efficient data analysis tools in many applications such as social network, electronic commerce, GIS, etc. Recently, semi-supervised clustering, for example, semi-supervised K-Means, semi-supervised DBSCAN, semi-supervised graph-based clustering (SSGC) etc., which uses side information to boost the performance of clustering process, has received a great deal of attention. Generally, there are two forms of side information: seed form (labeled data) and constraint form (must-link, cannot-link). By integrating information provided by the user or domain expert, the semi-supervised clustering can produce expected results of users. In fact, clustering results usually depend on side information provided, so different side information will produce different results. In some cases, the performance of clustering may decrease if the side information is not carefully chosen. This paper addresses the problem of choosing seeds for semi-supervised clustering, especially for graph based clustering by seeding (SSGC). The properly collected seeds can boost the quality of clustering and minimize the number of queries solicited from users. For this purpose, we propose an active learning algorithm (called SKMMM) for the seeds collection task, which identifies candidates to solicit users by using the K-Means and min-max algorithms. Experiments conducted on some real data sets from UCI and a real collected document data set show the effectiveness of our approach compared with other methods.

Keywords. Active Learning; Graph Based Method; K-Means, Semi-Supervised Clustering.

1. INTRODUCTION

Recently, semi-supervised clustering (seed based clustering or constraints based clustering) has received a great deal of attention in researcher communities [1, 2, 8, 13, 14, 15, 21, 25, 28]. The advantage of semi-supervised clustering consists in possibility to use a small set of side information to improve clustering results. There are two kinds of side information including constraints and seeds (see Figure 1). Constraints include must-link and cannot-link pairwise dependencies in which must-link constraint between two objects x and y means that x and y

[*] This paper is selected from the reports presented at the 12th National Conference on Fundamental and Applied Information Technology Research (FAIR'12), University of Sciences, Hue University, 07–08/06/2019.

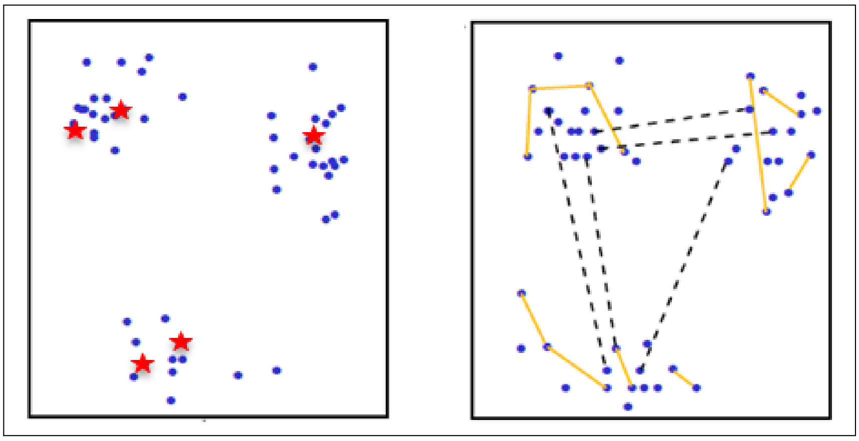

Figure 1. Two kinds of side information: (left) seeds are illustrated by red star points; (right) must-link and cannot-link constraints are respectively presented by solid and dash lines

should be grouped in the same cluster, and cannot-link constraint means that x and y should not be grouped in the same cluster. In the case of using seeds, a small set of labeled data will be provided from users/experts for semi-supervised clustering algorithms. In real applications, we hypothesize that the side information is available or can be collected from users.

Generally, semi-supervised clustering algorithms have two following important properties: (1) ability to integrate side information and (2) ability to boost the performance of clustering. Some principle techniques used in constraint based clustering include metric learning [9, 27], embedding constraints, kernel method, graph based method, etc. [13, 21]. In seed based clustering, a set of seeds can be used for initializing cluster centers in K-Means and Fuzzy C-Means [4], for automatically evaluating parameters in semi-supervised density-based clustering [10, 15], or identifying connected components for the partitioning process in semi-supervised graph based clustering (SSGC) [21]. The applications of semi-supervised clustering appear in many domains which include computer vision [8], Mining GPS Traces for Map Refinement [17], detecting fast transient radio anomalies [19], face grouping in video [26], deriving good partitioning that satisfies various forms of constraints in the k-anonymity model for privacy-preserving data publishing [8], and clustering medical publications for Evidence Based Medicine [16], etc.

In fact, seeds or constraints are randomly chosen for soliciting label from users. However, defining the label is a time consuming process, e.g., in speech recognition, annotating gene and disease [18], and the performance of clustering may decrease if the side information is not carefully chosen [15, 24]. The purpose of this paper is to develop an active learning method to collect seeds for semi-supervised graph based clustering. The active learning process is used along with semi-supervised clustering as shown in the Figure 2. Note that the active learning for semi-supervised classification has a long history but in a different context [18]. The seeds collected by our method can boost the performance of SSGC and minimize user queries compared with other methods. The idea of our method is to use a K-Means clustering algorithm in the first step and in the second step the min-max method will be used to select the candidates for getting labels from users. In summary, the contributions of this paper are

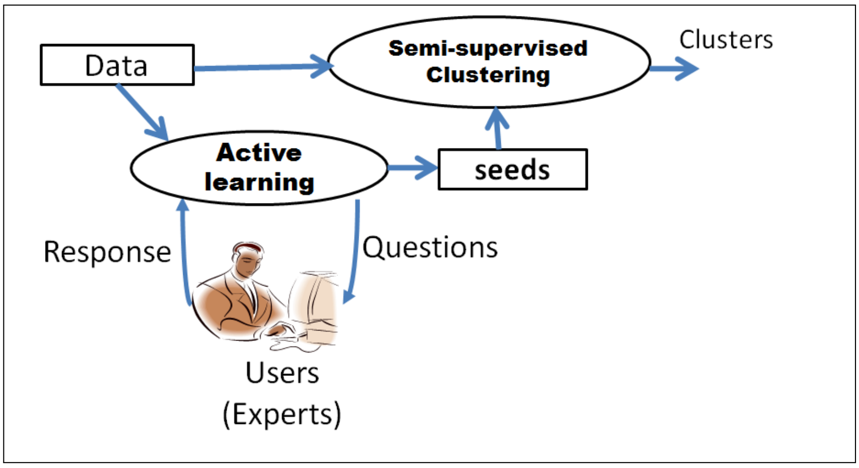

Figure 2. Relating between active learning and semi-supervised clustering

as follows:

- We survey some principle methods about seed based clustering and active seed selection methods for seed based clustering algorithms.
- We propose a simple but efficient method for collecting seeds applied for semi-supervised graph based clustering.
- We have conducted experiments for 8 data sets for comparing the proposed method with some reference methods. Moreover, we also create a Vietnamese document data set and propose to use it in an information extraction system. Finally, the effect of the parameter has also been analyzed for the proposed algorithm.

The rest of paper is organized as follows. Section 2 presents some related works. Section 3 introduces our new method for seeds collection. Section 4 describes the experiments that have been conducted on benchmark and real data sets. Finally, section 5 concludes the paper and discusses several lines of future researches.

2. RELATED WORK

2.1. Seed based clustering algorithms

As mentioned in the previous section, there are two kinds of semi-supervised clustering, in this section we focus on the seed based clustering. Generally, the seeds based clustering algorithms integrate a small set of seeds (labeled data points) in the process of clustering to improve clustering results. We will present some main works of seed based clustering hereafter.

In [15], a semi-supervised density based clustering algorithm named SSDBSCAN is presented. The SSDBSCAN extends the original DBSCAN algorithm by using a small set of labeled data to cope with the problem of finding clusters in distinct densities data. The objective of SSDBSCAN is to overcome this problem by using seeds to compute an adapted radius ϵ for each cluster. To do this, the data set is represented as a weighted undirected graph where each vertex corresponds to an unique data point and each edge between objects p and q has a weight defined by the $rDist()$ measure presented hereafter (see equation 1). The $rDist(p, q)$ measure illustrates the smallest radius value for which p and q are core points and directly density connected with respect to $MinPts$. Thus, $rDist()$ can be formalized as in the equation 1

$$rDist(p, q) = \max\{cDist(p), cDist(q), d(p, q)\}, \tag{1}$$

where $d()$ is the metric used in the clustering, $o \in X$ and $cDist(o)$ is the minimal radius such that o is a core-point and has MinPts nearest-neighbors.

Given a set of seeds D, the process of constructing clusters in SSDBSCAN is as follows. Using the distance $rDist()$, it is possible to construct a cluster C that contains the first seed point p, by first assigning p to C and then adding the next closest point in term of $rDist()$ measure to C. The process will continue until there is a point q having a different label from p. At that time, the algorithm backtracks to the point o that has the largest value of $rDist()$ before adding q. The current expansion process stops and includes all points up to but excluding o, having a cluster C containing p. Conceptually, this is the same as the constructing a minimum spanning tree (MST) in a complete graph where the set of vertices is equal X and the edge weights are given by $rDist()$. The complexity of SSDBSCAN is higher than that of DBSCAN, however, SSDBSCAN can detect the clusters in different densities.

In [21], the semi-supervised graph based clustering is proposed. In the algorithm, seeds are mainly used for helping in the partition step to form connected components. The SSGC includes two steps as follows:

Step 1: Graph partitioning by a threshold based on seeds: This step aims to partition a graph into connected components by using a threshold θ in a loop: all edges which have weight less than θ will be removed to form connected components at each step. The value of θ is assigned to 0 at first step and is incremented by 1 after each step. This loop will stop when the cut condition is satisfied as follows: each connected component contains at most one type of seeds. After finding the connected components, main clusters are constructed by propagating label in the obtained components.

Step 2: Detecting noises and constructing final clusters: The remaining points (graph nodes) that are not any main clusters will be put into two sets: Points that have edges assigned to related to one or more clusters and others points which can be considered as isolated points. In the first case, points are assigned to main clusters with the largest related weight. For the isolated points in the second case, two choices are possible depending on the users expectation: Either removing them as noises or labeling them.

In [7], the authors use some seeds to help the K-Means clustering in the step of finding k centers, named SSK-Means (see Algorithm 1). Although the proposed method is simple, however the clustering results are stable and SSK-Means overcome the effect of the choosing k centers at the initial step as the traditional K-Means algorithm.

In [13], the seed based on fuzzy C-Means is introduced. There seeds are used in the step of calculating the cluster memberships and object function to converge a good value.

Algorithm 1 The algorithm SSK-Means

```
Input: Data set $X = \{x_i\}_{i=1}^{N}$, number of clusters $K$, set of seeds $S = \{S_l\}_{l=1}^{k}$
Output: $k$ clusters of $X = \cup_{l=1}^{k} X_l$
Process:
Initializing: $\mu_h^{(0)} \leftarrow \frac{1}{|S_h|} \sum_{x \in S_h} x$, for $h = 1, ..., K; t \leftarrow 0$
repeat
    Assigning_cluster: Identify the cluster for $x$: $h^*$ (i.e. set $X_{h^*}^{(t+1)}$),
    $h^* = \text{argmin}\|x - \mu_h^{(t)}\|^2$
    estimating_means: $\mu_h^{(t+1)} \leftarrow \frac{1}{|X_h^{(t+1)}|} \sum_{x \in X_h^{(t+1)}} x$
    $t \leftarrow (t+1)$
until (Convergence)
Return $k$ clusters;
```

2.2. Active learning for semi-supervised clustering

While active learning algorithms for supervised classification has been investigated for a long period of time [18], the problem of active learning for semi-supervised clustering was mentioned for the first time in the research on integrating prior knowledge in clustering proposed in 2002 [14]. Recently a great number of research results on constraint clustering are reported. The principle idea is to select the most useful constraints/seeds so that they not only boost the clustering performance but also minimize the number of queries to the user. In [6], Basu et al proposed a method based on min-max strategy for constrained K-Means clustering. In [23], Vu et al proposed a graph-based method to collect constraints for any kind of semi-supervised clustering algorithms, in [2, 3], Abin et al. introduced a kernel-based sequential approach for active constraint selection for K-Means and DBSCAN.

In [22], a seed collection method based on min-max have been proposed, we refer as the SMM method. SMM collects the seeds based on the min-max strategy. The idea of SMM is to find a set of points which can cover the distribution of input data. Given a data set X, SMM uses an iterative approach to collect a set of seed candidates Y. At step 1, the y_1 is randomly chosen from X and $Y = \{y_1\}$, at step t $(t > 1)$, a new seed candidate y_t is selected and labeled by users/experts according to the equation 2:

$$y_t = \text{argmax}_{x \in X} \left(\min\{d(x, y_i)\}, i = 1 \ldots t-1\right) \tag{2}$$

where $d(.,.)$ denotes the distance defined in the space of the objects. After that, y_t will be added in the set Y.

The SMM has been shown to be efficient for semi-supervised K-Means clustering algorithm, the clustering based on partition. However, for the algorithms that can detect cluster with different densities, it does not work well. Figure 3 shows an example of the SMM method.

In [22], a graph based method for seeds collecting has been introduced (SkNN). The method uses a k-nearest neighbor graph to express the data set and each point in data set is assigned by a local density score using the graph. SkNN can collect seeds for any kind of

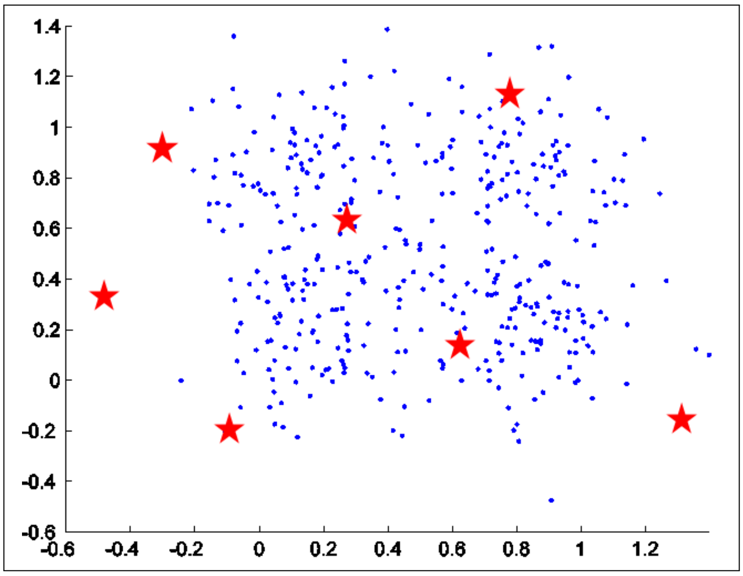

Figure 3. An example of seeds (red star points) collected by the min-max method

semi-supervised clustering algorithms. However, the complexity of the SkNN is $O(n^2)$.

3. THE PROPOSED METHOD

The SMM method is efficient when collecting seeds for the partition clustering, i.e. K-Means, Fuzzy C-Means, it does not work well for the semi-supervised clustering algorithms that produce clusters with arbitrary shapes such as SSDBSCAN, SSGC, etc. To overcome this limitation, in this section we propose a new algorithm combining the K-Means algorithm and the SMM method for Seed collecting problem, named SKMMM.

Given a data set X with n points, at first step, we use K-Means algorithm to partition X into c clusters. The number of clusters in this step is chosen big enough, i.e. up to $\sqrt{n}$ [29]. In the second step, we use the min-max method to choose seed candidates to get label from users. In the step 9, with the active learning context, we always assume that the users/expert can respond all the questions proposed by the system. The detail steps of the SKMMM algorithm are presented in Algorithm 2.

The complexity of the using K-Means algorithm is $O(n \times c)$, the complexity of the process of choosing an initial y_t seed that is nearest the center of an arbitrary cluster is $O(n/c)$, assume that, we need to collect v seeds, so the total complexity of the KMMFFQS is $O(n \times c) + O((v \times n)/c)$.

We also note that, in some recent works, the idea of using K-Means in the step of reducing the size of data set has been applied in many ways such as finding clusters with arbitrary shape [11], finding minimum spanning tree [29], and collecting constraints for semi-supervised clustering [20]. Figure 3 shows an example of the seed candidates selected by the SKMMM method. At first step, the data set will be partitioned by K-Means to form c small clusters. Based on the obtained clusters, the min-max strategy will identify the candidates to get label from users. By this way, the users question can significantly reduce and the

Algorithm 2 SKMMM

```
Input: A set of data points $X$, number of clusters $c$ for $K$-Means
Output: The collected seeds set $Y$
$Y = \emptyset$
Using $K$-Means for partitioning $X$ into $c$ clusters
Choosing an initial $y_1$ seed that is nearest the center of an arbitrary cluster
$Y = Y \cup \{label(y_1)\}$
$t = 1$
repeat
    $t = t + 1$
    Using min-max method to find the candidate cluster $c_t$; $y_t$ is chosen as the nearest
    point of the selected cluster
    Querying to users to get label for $y_t$
    $Y = Y \cup \{label(y_t)\}$
until user_stop = true
Return $Y$
```

process of collecting seeds do not depend on the shape of clusters.

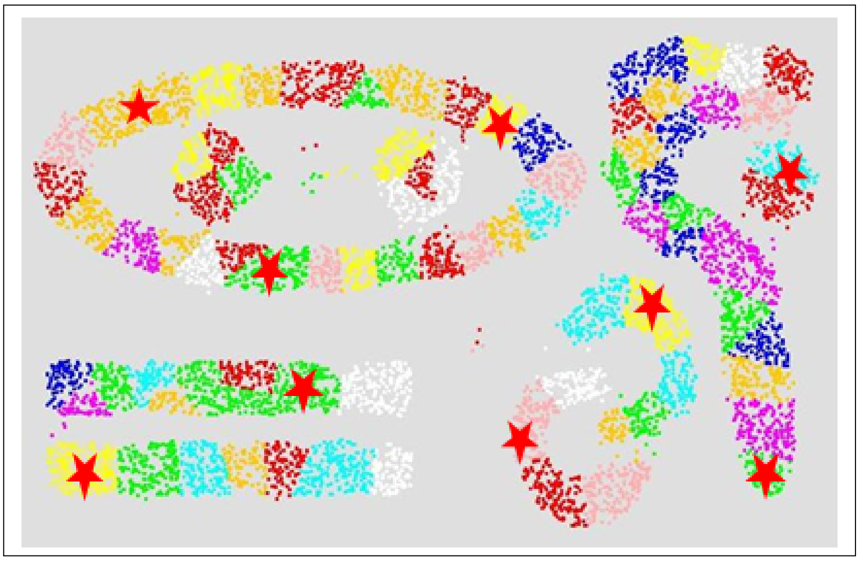

Figure 4. Example of seed candidates (red stars) selected by SKMMM method

4. EXPERIMENT RESULTS

4.1. Experiment setup

To evaluate our new algorithm, we have used 7 data sets from UCI machine learning [5] and one document data set collected from Vietnamese journal named $D1$. These UCI data sets have been chosen because they facilitate the reproducibility of the experiments and because some of them have already been used in semi-supervised clustering algorithms [8]. The details of these data sets are shown in Table 1 in which n, m, and k respectively are the number of data points, the number of attributes, and the number of clusters. The $D1$ data set consists 4000 documents in some topic such as sport, car, education, etc. getting from some Vietnamese journals. For the feature extraction of $D1$ data set, following the method presented in [12], the document is transformed into a vector using the TF-IDF method.

Table 1. Details of the data sets used in experiments

ID	Data	#Objects	#Attributes	#Clusters
1	Ecoli	336	7	8
2	Iris	150	4	3
3	Protein	115	20	6
4	Soybean	47	35	4
5	Zoo	101	16	7
6	Haberman	306	3	2
7	Yeast	1484	8	10
8	$D1$	4000	30	10

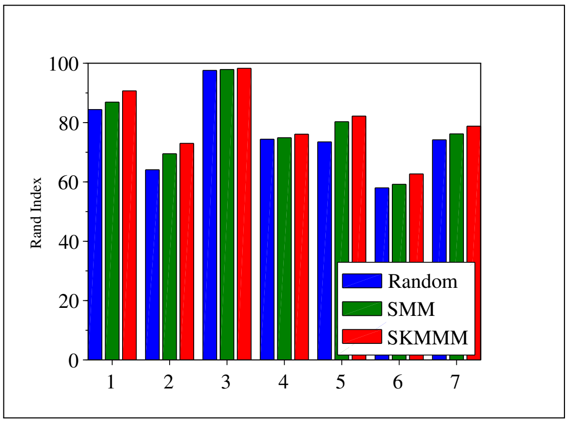

Figure 5. Rand Index measure for three seed collection methods with SSGC

To estimate the clustering efficiency we have used the Rand Index (RI) measure, which is widely used for this purpose in different researches [8]. The RI calculates the matching between the true partition (P_1) and the obtained partition (P_2) of each data set by the evaluated clustering algorithm. To compare two partitions P_1 and P_2, let u be the number of decisions where x_i and x_j are in the same cluster in both P_1 and P_2. Let v be the number of decisions, where the two points are put in different clusters in both P_1 and P_2. The RI measure is calculated using the following equation

$$RI(P_1, P_2) = \frac{2(u+v)}{n(n-1)}. \tag{3}$$

The value of RI is in the interval [0..1]; $RI = 1$ when the clustering result corresponds to the ground truth or user expectation. The higher the RI, the better the result.

4.2. Experiment results

In this section, we present clustering results by using the SKMMM, SMM and the random method. Two aspects examined to get labels are: The quality of clustering and the number of queries needed. We note that, for each method, the process of collecting seeds will stop when at least one seed had chosen for each cluster.

Figure 5 shows the results of clustering using the seeds of three methods, respectively. It can be seen from the figure that using seeds collected by SKMMM, the quality of SSGC clustering is better than using seeds collected by other methods. It can be explained by the fact that when we use the K-Means algorithm at the first step, the number of candidate is decrease and the seeds collected by the SKMMM are all near the center of each cluster so it is good for the propagation process in the SSGC algorithm. With the random method, the candidates are randomly chosen to get label from users. So, the results is always not stable.

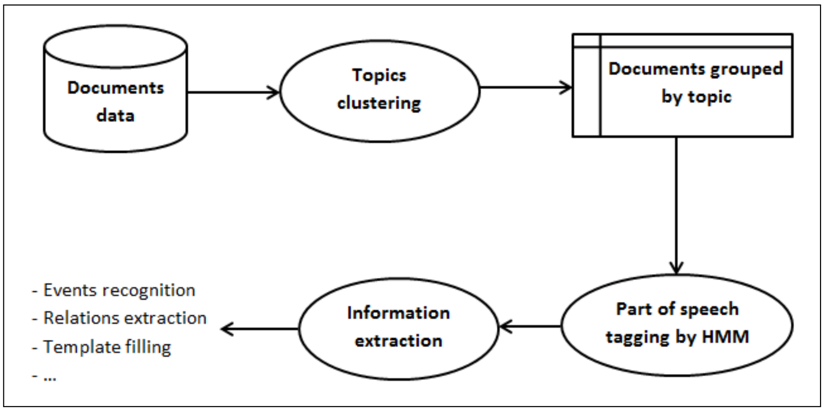

Figure 6. Information extraction schema

For the $D1$ data set, that is the document data set collected from Vietnamese journals, as we can see, documents now have been increasing very fast and hence the problem of document mining/processing is a crucial task. Clustering algorithms can be used to discover latent concepts in sets of unstructured text documents, and to summarize and label such collections. In the information extraction problem, in which we need to extract information from a large of document data sets, we can apply the clustering to partition documents in topics and after that each topic will be used for the information extraction task. This idea of using the clustering step can reduce the size of data set that we want to analysis when working with large scale data sets. For real applications, we propose to use a clustering algorithm in a schema of an information extraction system as illustrated in the Figure 6.

Figure 7 illustrates the number of queries used in active learning process with three methods. As shown in the figure, the SKMMM method needs fewer queries than the SMM and the random method. This is a significant advantage for the actual problem when getting label takes a lot of time and effort [18]. We can explain by the fact that using the K-Means at the first step is a good way for sampling the data.

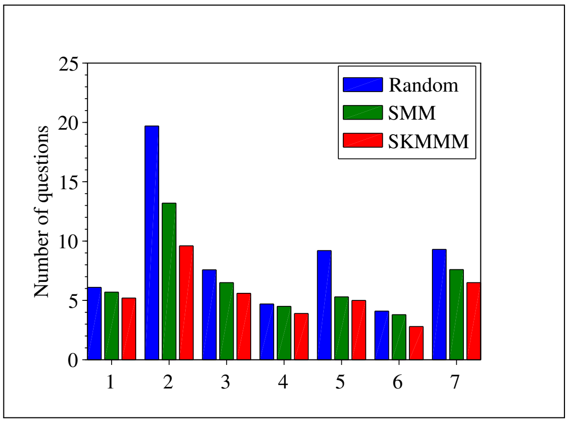

Figure 7. Number of queries for three methods with 8 data sets

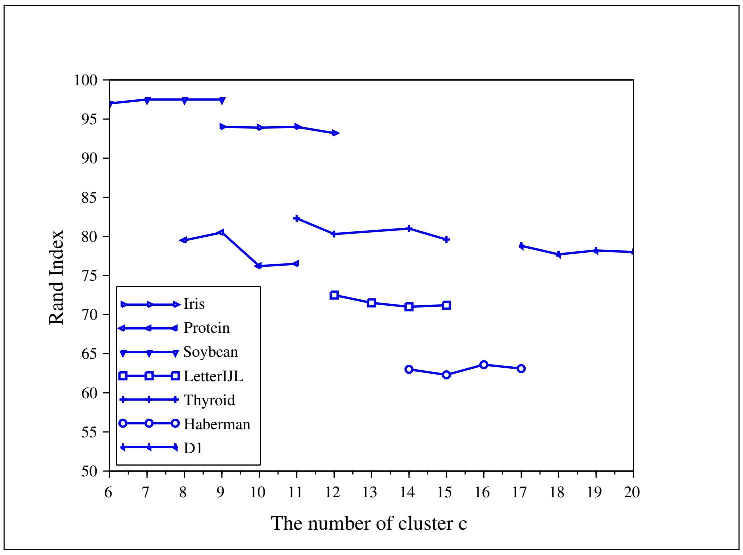

Figure 8. Clustering results of 7 data sets with some values of c

Another experiment that we did is to analysis the effect of the number of clusters c in the query selection step to the quality of clustering. Figure 8 shows the results of clustering for 7 data sets; in this experiment we choose the value of c is in the range of $[\sqrt{n}-3, ..., \sqrt{n}]$. We can see from these results, the best results for each data set can be obtained with c selected in these intervals. We also note that, the process of reducing candidates to get label from data do not depends on the size, shape, and density of clusters.

5. CONCLUSIONS

Active learning for semi-supervised clustering has received a lot of attention in the past two decades. In this paper, we present a new SKMMM algorithm for collecting seeds applied for the semi-supervised graph based clustering. Experiments conducted on some UCI data sets and a real collected document data set show that the proposed method is an efficient method which not only increases the clustering quality of SSGC algorithm but also decreases the number of user queries compared with several reference methods. In future work, we aim to develop other seed collection algorithms together with their practical applications and also to test with other kinds of semi-supervised clustering.

ACKNOWLEDGMENT

This research is funded by Vietnam National University, Hanoi (VNU) under project number QG.18.40.

REFERENCES

[1] A. Abin, "Clustering with side information: Further efforts to improve efficiency," *Pattern Recognition Letters*, vol. 84, pp. 252–258, 2016.

[2] A. Abin and H. Beigy, "Active selection of clustering constraints: a sequential approach," *Pattern Recognition*, vol. 47, no. 3, pp. 1443–1458, 2014.

[3] ——, "Active constrained fuzzy clustering: A multiple kernels learning approach," *Pattern Recognition*, vol. 48, no. 3, pp. 953–967, 2015.

[4] V. Antoine, N. Labroche, and V.-V. Vu, "Evidential seed-based semi-supervised clustering," in *In Proc. of the 7th Intl. Conf. on Soft Comput. and Intl. Syst. and 15th Int. Symp. on Adv. Intell. Syst.*, 2014, pp. 706–711.

[5] A. Asuncion and D. Newman, "Uci machine learning repository," in *University of California, Irvine, School of Information and Computer Sciences*, 2015. [Online]. Available: http://www.ics.uci.edu

[6] S. Basu, A. Banerjee, and R. J. Mooney, "Active semi-supervision for pairwise constrained clustering," in *Proceedings of the SIAM International Conference on Data Mining, SDM-2004*, 2004, pp. 333–344.

[7] S. Basu, A. Banerjee, and R. Mooney, "Semi-supervised clustering by seeding," in *In Proc. of 19th Intl. Conf. on Machine Learning*, 2002, pp. 281–304.

[8] S. Basu, I. Davidson, and K. L. Wagstaff, *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, 1st ed. Chapman and Hall/CRC Data Mining and Knowledge Discovery Series, 2008.

[9] M. Bilenko, S. Basu, and R. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *In Proc. 21st Int. Conf. on Machine Learning*, 2004. [Online]. Available: doi⟩10.1145/1015330.1015360

[10] C. Bohn and C. Plant, "Hissclu: a hierrachical density-based method for semi-supervised," in *Intl. Conf. on Extending Database Technology*, 2008, pp. 440–451. [Online]. Available: doi⟩10.1145/1353343.1353398

[11] V. Chaoji, M. A. Hasan, S. Salem, and M. J. Zaki, "Sparcl: an effective and efficient algorithm for mining arbitrary shape-based clusters," *Knowl. Inf. Syst.*, vol. 21, no. 2, pp. 201–229, 2009.

[12] I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine Learning*, vol. 42, pp. 143–175, 2001.

[13] N. Grira, M. Crucianu, and N. Boujemaa, “Active semi-supervised fuzzy clustering,” *Pattern Recognition*, vol. 41, no. 5, pp. 1834–1844, 2008.

[14] 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 *Proceedings of the Nineteenth International Conference*, 2002, pp. 307–314. [Online]. Available: http://ilpubs.stanford.edu:8090/528/

[15] L. Lelis and J. Sander, “Semi-supervised density-based clustering,” in *2009 Ninth IEEE International Conference on Data Mining*, 2009, pp. 842–847. [Online]. Available: DOI:10.1109/ICDM.2009.143

[16] A. S. Saha, D. Molla, and K. Nandan, “Semi-supervised clustering of medical text,” in *Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP)*, 2016, pp. 23–31. [Online]. Available: https://www.aclweb.org/anthology/W16-4205.pdf

[17] S. Schrodl, K. Wagstaff, S. Rogers, P. Langley, and C. Wilson, “Mining gps traces for map refinement,” *Data Min. Knowl. Discov.*, vol. 9, no. 1, pp. 59–87, 2004.

[18] B. Settles, “Active learning literature survey,” in *Computer Sciences Technical Report 1648. University of WisconsinMadison*, 2010. [Online]. Available: http://digital.library.wisc.edu/1793/60660

[19] D. Thompson, W. Majid, C. Reed, and K.-L. Wagstaff, “Semi-supervised eigenbasis novelty detection,” *Statistical Analysis and Data Mining*, vol. 6, no. 3, pp. 195–204, 2013.

[20] H. B. Toon van Craenendonck, S. Dumancic, “COBRA: A fast and simple method for active clustering with pairwise constraints,” in *Proc. of IJCAI*, 2017, pp. 2871–2877. [Online]. Available: arXiv.org⟩cs⟩arXiv:1801.09955

[21] V.-V. Vu, “An efficient semi-supervised graph based clustering,” *Intelligent Data Analysis.*, vol. 22, no. 2, pp. 297–307, 2018.

[22] V.-V. Vu and N. Labroche, “Active seed selection for constrained clustering,” *Intelligent Data Analysis*, vol. 21, no. 3, pp. 537–552, 2017.

[23] V.-V. Vu, N. Labroche, and B. Bouchon-Meunier, “Improving constrained clustering with active query selection,” *Pattern Recognition*, vol. 45, no. 4, pp. 1749–1758, 2012.

[24] K. Wagstaff, S. Basu, and I. Davidson, “When is constrained clustering beneficial, and why?” in *In AAAI*, 2006. [Online]. Available: http://new.aaai.org/Papers/AAAI/2006/AAAI06-384.pdf

[25] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrodl, “Constrained k-means clustering with background knowledge,” in *Proc. of Intl. Conf. on Machine Learning*, 2001, pp. 577–584.

[26] B. Wu, Y. Zhang, B.-G. Hu, and Q. Ji, “Constrained clustering and its application to face clustering in videos,” in *Proc. of CVPR*, 2013, pp. 3507–3514.

[27] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. J. Russell, “Distance metric learning with application to clustering with side-information,” in *NIPS*, 2002, pp. 505–512. [Online]. Available: http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information.pdf

[28] R. Yan, J. Zhang, J. Yang, and A. Hauptmann, “A discriminative learning framework with pairwise constraints for video object classification,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 4, pp. 578–593, 2004.

[29] C. Zhong, M. Malinen, D. Miao, and P. Franti, “A fast minimum spanning tree algorithm based on k-means,” *Inf. Sci.*, vol. 295, pp. 1–17, 2015.

Received on August 07, 2019
Revised on September 26, 2019