ACTIVE CONSTRAINTS SELECTION BASED ON DENSITY PEAK ESTIMATION

Pham Gia Bao[1], Dinh Quoc Viet[1], Le Tuan Linh[1], Phung Van Linh[1], Ha Khanh[1], Vu Viet Vu[1], Tran Doan Vinh[2], Vu Viet Thang[3]

Received: 13 May 2025/ Accepted: 15 August 2025/ Published: 25 August 2025

Https://doi.org/10.70117/hdujs.E10.08.2025.907

Abstract: *Semi-supervised clustering, which integrates side information from users to enhance clustering performance, has gained considerable attention in the research community. However, the quality of clustering is highly dependent on the side information provided, and different inputs can lead to different results. In this paper, we propose an active learning approach for selection good constraints, which employs a min-max strategy and density-based estimation of data points to optimize the constraints selection process. Experimental evaluations on datasets from UCI and an real face image data show the effectiveness of our method.*

Keywords: *Clustering, semi-supervised clustering, constraint, density peak, active learning, min-max method.*

1. Introduction

In recent years, semi-supervised clustering has gained significant attention within the research community [1-2]. Its primary advantage is the ability to improve clustering performance by leveraging a small amount of side information. This supplementary information falls into two main categories: constraints and seeds. Constraints, usually expressed as must-link or cannot-link, show the relationship bewtween two data points that shoul be in the same cluster or not.

In practical applications, it is assumed that such side information is either readily available or can be obtained from users or from domain experts. One of the first research about semi-supervised clustering with constraints was proposed by Wagstaff in 2000 [2].

When integrating constraints into semi-supervised clustering, a key challenge is determining how to obtain a high-quality set of constraints before the clustering process begins. While extensive research has been conducted on constraint-based clustering [3], most approaches assume that users passively provide well-chosen constraints to guide the algorithm. A more effective alternative is to adopt an active learning approach, where users are actively involved in selecting constraints. However, selecting constraints strategically is crucial, as poorly chosen constraints can lead to suboptimal algorithm convergence [1,4]. Additionally, with a dataset containing n data points, the number of possible must-link (ML) or cannot-link (CL) constraints can reach $(n \times (n-1))/2$, making exhaustive selection impractical. This challenge falls under the broader field of active learning [3], which aims to optimize constraint selection for improved clustering accuracy and efficiency.

[1] CMC University, Ha Noi; Email: vvvu@cmcu.edu.vn
[2] VNU University of Education, Vietnam National University, Hanoi
[3] Hanoi University of Industry, Hanoi

Figure 1 shows an example of constraints. In the figure must-link constraints are linked by the solid line while connot-link constraints are linked in the dash lines.

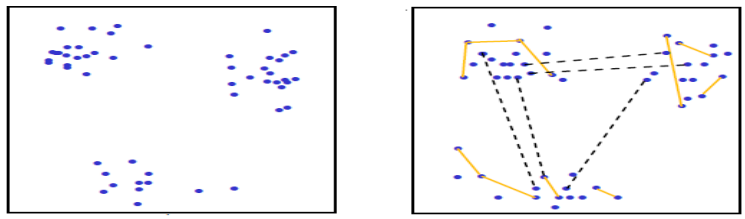

Figure 1. An example of data set (left) and constraints (right)

In this paper, we introduce a novel active learning approach for gathering constraints from users or domain experts, the work is extended from paper in [3]. Our proposed method is built upon the principles of density peak estimation and the min-max strategy to optimize constraint selection. Experimetation conducted on some real data set illustrates the effectiveness of our method.

The rest of the paper is organized as follows: The material and method will be presented in the section 2, the experimentation and discussion will be showed in section 3. The section 4 concludes the papers.

2. Material and method

2.1. Material

We use six data sets from UCI and one data set collected by our group (named FaceCMC) which consists of faces extracted from camera in some students class. The features of FaceCMC are extracted by using an pre-trained model named Deepface. We use the MCSSGC [8], a semi-supervised graph-based clustering algorithm, to measure the effectiveness of constraints collected by our algorithm.

Table 1.Main characteristics of the datasets using in the experimentations

No	Data	n	m	k
1	Iris	150	4	3
2	Soybean	214	9	6
3	Ecoli	336	8	8
4	Protein	115	20	6
5	Zoo	101	16	7
6	Thyroid	215	5	3
7	FaceCMC	6000	512	14

Given P_1, P_2 are two partitions, the Rand Index [9] score is used for evaluation of clustering results as the following equation:

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

The 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. In our experimentations we use the Rand Index in percentage.

2.2. Density peak clustering

The Density Peak Clustering (DPC) is proposed in 2014 that is the clustering method based on density concept [10]. The key idea of DPS is based on two criteria: (1) points with high local density should be cluster centers and (2) cluster centers should be far from other high-density points to ensure well-separated clusters.

Firstly, the local density of each data point x_i is calculted as in the equation (1):

$$\rho_i = \sum_j \mathrm{X}\left(d_{ij} - d_c\right) \quad (1)$$

in which X(x) = 1 if X < 0 and X(x) = 0 otherwise, d_c is an input parameter.

Secondly, the distance $\delta(x_i)$ is measured as the minimum distance between the point i and any other point with higher density as shown in equation (2).

$$\delta_i = \min_{j:\rho_j > \rho_i} \left(d_{ij}\right) \quad (2)$$

Using $\rho(x_i)$ and $\delta(x_i)$, a decision graph will be build in which the x-axis and the y-axis are respectively the rho and delta values for whole data set. An example of decision graph is illustrated in figure 2. From the decision graph, we can identify peaks as points with maxima local density (at the right up corner). Some improved version of DPC can be cited here such as these works in [10-13]. DPC is particularly useful for identifying clusters of arbitrary shapes and handling uneven densities, making it an effective method for unsupervised learning in diverse applications.

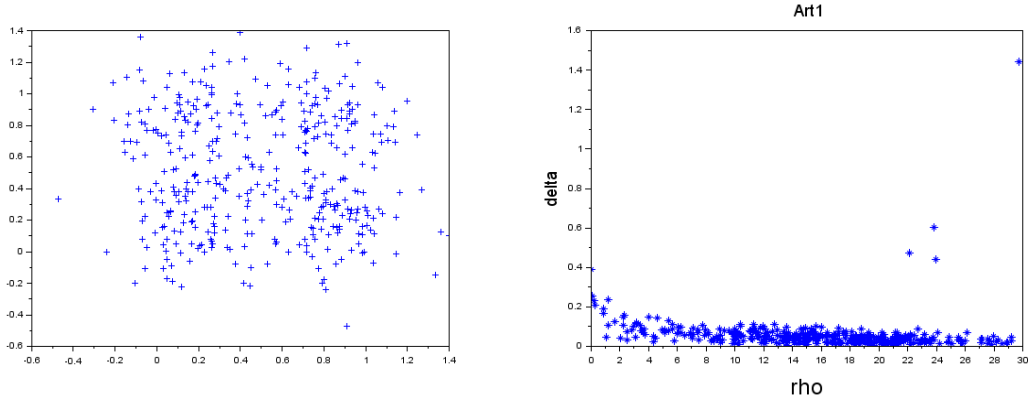

Figure 2. An example about data (left) and its decision graph (right)

2.3. Active learning method for semi-supervised clustering

The idea of employing active learning in semi-supervised clustering was first introduced in a study that explored the integration of prior knowledge into clustering algorithms [14]. Unlike traditional clustering methods, which operate purely in an unsupervised manner, semi-supervised clustering benefits from additional user-provided information, making it possible to enhance clustering performance and produce more

meaningful results. Instead of relying on a large, randomly labeled dataset, an active learning model intelligently queries an oracle (e.g., a human annotator or expert system) to obtain labels for carefully chosen instances (see figure 3). This approach is highly beneficial, particularly in modern machine learning applications where unlabeled data is abundant and easy to acquire, but labelling is often difficult, time-consuming, and expensive. The challenge in applying active learning to semi-supervised clustering lies on the selection of informative constraints. Unlike classification tasks, where the learner queries labels for individual data points, semi-supervised clustering requires querying relationships between points to define must-link and cannot-link constraints effectively. If poorly chosen constraints are introduced, they can mislead the clustering process and result in suboptimal partitions of data. Given that a dataset with n data points could have up to (n × (n-1))/2 possible must-link or cannot-link constraints, choosing the most beneficial constraints is a critical challenge.

Current research in this field aims to refine selection strategies for active learning in semi-supervised classification, ensuring that queries yield the most significant improvements in model performance. Several approaches, such as uncertainty sampling, density-weighted selection, and entropy-based methods, have been explored to optimize query efficiency. Some works have done for these topics such as in [3-6].

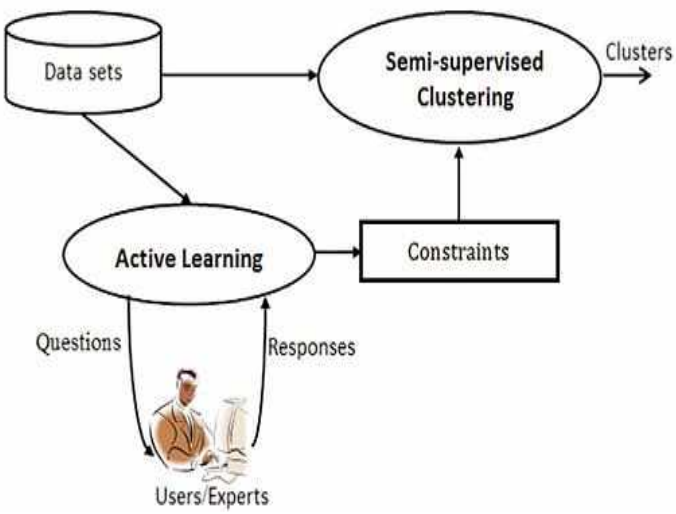

Figure 4. The schema of active learning for semi-supervised clustering [18]

2.4. The min-max method

Given a data set X with n point, the min-max method uses an iterative approach to build a subset Y as the following steps:

- Step 1: Y= $\{y_1\}$, y_1 is randomly chosen from X
- Step 2: For step t ($t \leq T$), y_t is identified as follow:

$$y_t = \text{argmax}_{x \in X}(\min\{d(x; y_i)\}); i = 1,\ldots t\text{-}1; Y = Y \cup \{y_t\};$$

An example of min-max method is shown in figure 5. It can be seen from the figure that the points collected by min-max (star points) can cover whole data set.

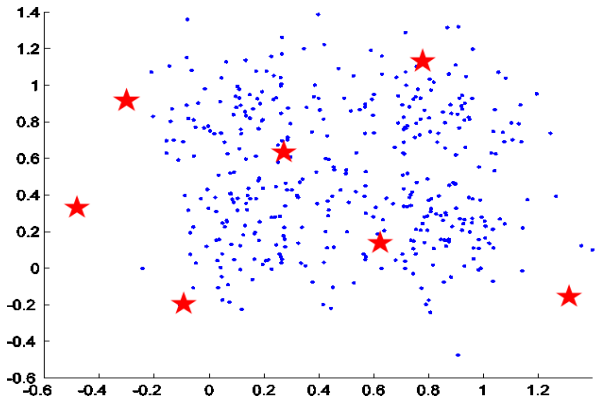

Figure 5. An example of min-max method: star points are collected by min-max method.

3. The proposed method

Using the local estimation data, decision graph in DPC and the min-max method, we have proposed an active constraints selection method as presented in Algorithm 1.

Algorithm 1: Active constraint selection based on density peaks;
Input: a data set X = {x_1, x_2, …, x_n}, d_c, v, c;
Output: a set of v constraints
Begin
Step 1: Calculate local density $\rho(x_i)$ for each data point
Step 2: Calculate $\delta(x_i)$ for each data point
Step 3: Creat the decision graph using $\rho(x_i)$ and $\delta(x_i)$
Step 4: Select skeleton peaks from the decision graph
Step 5: P = [p_1, p_2,…,p_k]; stt = 1; V ={};
Step 6: L= {c% lowest density points of X}
Step 7: *Repeat*
Step 8: Select the point x in L that follows the min-max method;
Step 9: t = 1;
Step 10: *Repeat*
Step 11: Question to users for getting label of (P_t, x);
Step 12: t = t +1; V = V {(P_t, x)}; stt = stt+1;
Step 13: *Until* (label(P_t, x) = CL) or (stt == v);
Step 14: k = k+1; P_k = x;
Step 15: *Until* (user_stop = true) or (stt ==v);
Step 16: Output the set of constraints collected;
End

In steps 1-3, the decision graph has been built as in DPC method. At step 4, the initial skeleton will be chosen that can be seen as basic set before applying min-max method. Steps 7-15, this is an loop process for selecting constraints: the point with farthest distance to Y and appearing in the top c% of lowest density points in data will be chosen to get label from users. In fact, the min-max method is applied in some problems such as finding the k centers for K-Means, collecting k seeds for seed based active learning clustering, etc.

In general, the complexity of Algorithm 1 is $O(n^2)$. However, for low dimentional data, we can use some data structure such as R-Tree that can reduce the complexity of the algorithm as O(nlogn).

4. Results and discussion

To measure the performance of the proposed algorithm, we use a semi-supervised clustering algorithm named MCSSGC which is published in [8]. We will compared the results obtaned by MCSSGC using constraints collected by algorithm 1 and constraints randomly chosen. The results have shonw in the table 2. It can be seen from results, MCSSGC using constrains collected by our algorithm obtained the better results compared with the random method. We can explain by the fact that by combining min-max method and local density estimation measure, we can choose the good candidate to get label from users hence help the clustering process in fiding clusters. The visualisation of the art1 data set and constraints collected by Algorithm 1 is also presented in the figure 6. As explained, we can see the constraints can cover points in the lowest density of data that is the *hard* points for clustering.

Table 2. Results of rand index measure (%)

No	Data	Number of constraints	MCSSGC+ Random	MCSSGC+ proposed
1	Iris	140	91.4	93.7
2	Soybean	60	98.9	100
3	Ecoli	200	90.1	91.5
4	Protein	120	81.9	83.3
5	Zoo	100	98.3	98.8
6	Thyroid	180	80.9	81.3
7	FaceCMCU	600	85.6	88.9

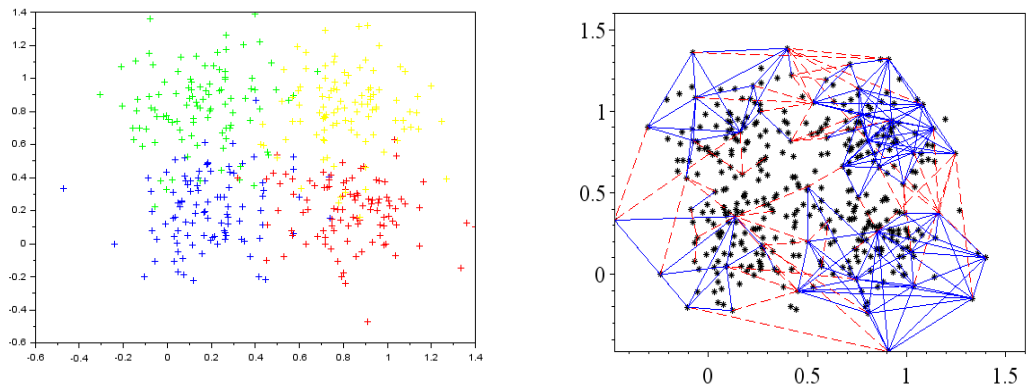

Figure 6. The visualisation of art1 data set and The constraints collected by our method: must-link and cannot-link constraints are illutrated by solid lines and dash linea

5. Conclusion

This paper has introduced an efficient method for constraints collection for semi-supervised clustering algorithms. The key idea is that each data point is estimated using its local density score, and then the min-max method will be applied to detect the interesting points to form constraints. Results obtained when using some real data sets from UCI and an face data set show the effectiveness of the proposed method.

References

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

[2] Kiri Wagstaff, Claire Cardie (2000), *Clustering with Instance-level Constraints*. ICML: 1103-1110.

[3] Viet-Vu Vu et al. (2022), *Active constraints selection based on density peak*. Proc. of International Conference on Advanced Communications Technology, 447-452.

[4] Nizar Grira, Michel Crucianu, Nozha Boujemaa (2008), *Active semi-supervised fuzzy clustering*. Pattern Recognit. 41(5): 1834-1844.

[5] Ahmad Ali Abin, Viet-Vu Vu (2020), *A density-based approach for querying informative constraints for clustering*. Expert Syst. Appl. 161: 113690.

[6] Viet-Vu Vu, Nicolas Labroche, Bernadette Bouchon-Meunier (2012), *Improving constrained clustering with active query selection.* Pattern Recognit. 45(4): 1749-1758.

[7] Sugato Basu, Arindam Banerjee, Raymond J. Mooney (2004), *Active Semi-Supervision for Pairwise Constrained Clustering*. SDM, 333-344.

[8] Viet-Vu Vu, Hong-Quan Do, Vu-Tuan Dang, Nang-Toan Do (2019), *An efficient density-based clustering with side information and active learning: A case study for facial expression recognition task*. Intell. Data Anal. 23(1): 227-240.

[9] W. M. Rand (1971), *Objective criteria for the evaluation of clustering methods*. Journal of the American Statistical Association. American Statistical Association. 66, 846–850.

[10] Rodriguez, A.; Laio (2014), *A. Clustering by fast search and find of density peaks*. Science, 344, 1492.

[11] Mehmood, R.; Zhang, G.; Bie, R.; Dawood, H.; Ahmad, H. (2016), *Clustering by fast search and find of density peaks via heat diusion*. Neurocomputing, 208, 210-217.

[12] Wang, S.; Wang, D.; Li, C.; Li, Y.; Ding, G. (2016), *Clustering by Fast Search and Find of Density Peaks with Data Field*. Chin. J. Electron, 25, 397-402.

[13] Du, M.; Ding, S.; Jia, H. (2016), *Study on density peaks clustering based on k-nearest neighbors and principal component analysis*. Knowl.-Based Syst., 99, 135-145.

[14] Burr Settles (2009), *Active Learning Literature Survey*. Computer Sciences Tech nical Report 1648, University of Wisconsin-Madison.