Semi-Supervised Clustering Algorithms Through Active Constraints

Abdulwahab Ali Almazroi, Walid Atwa
Department of Information Technology-College of Computing and Information Technology at Khulais,
University of Jeddah, Jeddah, Saudi Arabia

***Abstract*—Pairwise constraints improve clustering performance in constraint-based clustering issues, especially since they are applicable. However, randomly choosing these constraints may be adverse and minimize accuracy. To address the problem of random choosing pairwise constraints, an active learning method is used to identify the most informative constraints, which are then selected by the active learning technique. In this research, we replaced random selection with an active learning strategy. We provide a semi-supervised selective affinity propagation clustering approach with active constraints, which combines the affinity propagation (AP) clustering algorithm with prior information to improve semi-supervised clustering performance. Based on the neighborhood concept, we select the most informative constraints where neighborhoods include labelled examples of various clusters. The experimental results on eight real datasets demonstrate that the proposed method in this paper outperforms other baseline methods and that it can improve clustering performance significantly.**

Keywords—Semi-supervised; pairwise constraints; affinity propagation; active learning

I. Introduction

In data mining, clustering is an unsupervised learning technique that divides a data collection into k clusters based on how similar or dissimilar data examples are within a cluster and outside of it. Clustering with restrictions, or semi-supervised clustering, has drawn a lot of attention from researchers in the past few years. By utilizing user-provided side information, semi-supervised clustering seeks to enhance clustering performance. Pairwise restrictions, or must-links (*ML*) and cannot-links (*CL*), are the most often utilized information in semi-supervised clustering. Instances x_i and x_j must be assigned to the same cluster according to the constraint $ML(x_i, x_j)$, but a $CL(x_i, x_j)$ specifies that they must be assigned to separate clusters [1, 2].

Constraints have been shown in several earlier research to improve clustering performance. However, incorrect constraint selection can also degrade the clustering performance [3-5]. Furthermore, getting pairwise constraints usually necessitates a user to examine the relevant data points by hand, which can be expensive and time-consuming. Most semi-supervised clustering techniques already in use choose all of their constraints at random. Therefore, these methods are unable to predict the impact of a particular constraint on the algorithm [6-8].

The affinity propagation (AP) method is a highly effective clustering method for data mining. Compared to standard clustering methods, the AP method is capable of clustering large-scale and multi-cluster datasets quickly. Furthermore, the AP method does not need to predetermine the initial centres of the cluster and cluster number, which allows it to avoid getting locked in the local optimal setting [15]. Thus, utilizing the AP technique is preferable since the clustering algorithm may provide higher-quality component clusters [9].

In this paper, we maximize the pairwise constraint selection for semi-supervised clustering by combining the affinity propagation (AP) clustering algorithm with prior information based on the neighborhood notion. A neighborhood is a collection of data items that must-link constraints have determined to belong to the same class. It is well known that distinct neighborhoods belong to different classes since they are connected by cannot-link constraints. Our goal is to choose the most educational point to incorporate into the neighborhoods. After a point is chosen, its neighborhood is ascertained by querying the chosen point against the list of neighborhoods. Utilizing the neighborhood ideas has the major benefit of allowing us to obtain constraints by utilizing the neighborhood knowledge.

On UCI datasets, extensive research has been done with the MPCK-means semi-supervised clustering technique. According to experimental results, MPCK-means performs better when subjected to AML and ACL restrictions than when subjected to random selection constraints.

The rest of the paper is structured as follows. A concise overview of relevant research on active learning techniques is given in Section II. We present our suggested active learning algorithm in Section III. Section IV presents the outcomes of the experiment. In Section V, we finally wrap up the paper and talk about future directions.

II. Related Work

Semi-supervised clustering algorithms are proposed in recent years [10]. These algorithms are an extension of known unsupervised clustering algorithms [11, 12]. The methods utilized constraints in adapted clustering procedure for learning similarity metrics. In recent years, various constraint-based methods were proposed for clustering, specifically in clustering algorithms like spectral clustering and K-means [11].

Basu et al. proposed a pairwise constrained clustering framework and a new method for selecting information pairwise constraints for enhance clustering performance [13]. These two methods, as indicated by the authors, can handle large and high dimensional datasets. The result shows an improvement in clustering accuracy with little supervision required.

Shental et al. proposed a framework for the composition of side information in the form of equivalence constraints into the model estimation procedure [14]. The authors further introduce EM procedure and generalized EM procedure that handles both positive constraints (for the former), and negative constraints (for the latter). The algorithm shows significant improvements. In another study by Bilenko et al., a new clustering algorithm was proposed that integrates two methods, which are the constraint-based method and distance-function learning methods for semi-supervised clustering [15]. Based on an experimental study, the result revealed that the proposed algorithm provides better clusters.

In a study by Wagstaff et al., a constrained *k*-means clustering algorithm was proposed with background knowledge [11]. The experiment was conducted with artificial constraints on various datasets, the result shows a significant improvement. Rangapuram and Hein proposed a clustering method that is based on tight relaxation of constraint normalized cut [12]. The proposed method guarantees the satisfy all constrained. The method further allows the optimization of the trade-off between the number of violated constraints and normalized cut. The result shows some improvements.

In the last decades, active learning has been studied for supervised classification problem. Xiong et al., introduced a method that incorporates a neighborhood concept. Hence, each neighborhood is composed of labeled examples of distinct clusters based on pairwise constraints [16]. An evaluation of benchmark datasets shows that the proposed method outperforms the existing state-of-the-art.

Fernandes et al. proposed four active learning strategies for an evolutionary constrained clustering algorithm coined FIECE-EM. The proposed strategies utilizes key information from multitudes of sources like partition, population, and so on [17]. An empirical evaluation result shows that the proposed strategies attain a better result in comparison to various state-of-the-art. Based on the knowledge that choosing a constraint is critical because choosing it improperly may result in low clustering precision, a new active query mechanism was proposed by Kumar et al. The proposed mechanism selects queries by utilizing min-max criterion. Hence, the authors specifically focused on constraints selection to enhance clustering performance. The experimental result indicates that the proposed outperforms the existing state-of-the-art [18].

In another study by Nguyen and Smeulders, an algorithm was developed to construct classifier in a group of cluster representatives and further propagates the conducted classification to other samples through a local noise model [19]. The developed algorithm initially selects the most active samples for the avoidance of repeatable samples labels. In the active learning process, the algorithm adjusts the clustering by utilizing coarse-to-fine strategy. This is purposely to balance between large clusters merit and data representation accuracy. The result demonstrates the performance of the proposed algorithm.

Another study by Vu et al. proposed an efficient algorithm for active seeds selection [20]. The proposed algorithm depends on min-max approach which permits the coverage of large dataset and the selection of useful user queries. The result shows that the proposed algorithm performs very well. A semi-supervised clustering algorithm with a new method for selecting information instance-level constraints was proposed to enhance clustering accuracy [21]. The proposed algorithm is coined Constrained DBSCAN. The algorithm is aimed at selecting informative document pairs to retrieve user feedback. Hence, the authors used two kinds of instance-level constraints, which are cannot-link and must-link. For the former, it means that document pairs must always be placed in distinct groups, while for the latter, the document pairs must be in the same cluster. The result shows that a good clustering performance was achieved.

Wang and Davidson proposed a spectral clustering algorithm with active learning and further investigates active learning [22]. The authors also allow for the utilization of cannot-link and must-link constraints in the proposed algorithm. However, in distinction, their constraints are identified incrementally through oracle querying. Hence, the outline advantages of their proposed algorithm are the process of constraints querying that reduces error, and the combination of both soft and hard constraints. The results based on an experiment on existing benchmark show that the proposed algorithm outperforms existing baseline approaches [23].

Although many studies have investigated active constraints, there is limited research on leveraging existing information to identify the most informative constraints in ensemble clustering or on integrating active constraints with selective ensemble clustering results [24-26]. To address this gap, this paper introduces a semi-supervised selective affinity propagation clustering approach that incorporates active constraints, aiming to enhance the performance of semi-supervised clustering.

III. Active Affinity Propagation

Affinity propagation (AP) [9] is a clustering technique that groups data points into clusters according to their similarities. Messages are sent between data points iteratively via affinity propagation. These messages show how each data point is ideally suited to serve as a cluster center for additional data points. High affinity data points eventually become cluster centers, to which other points are allocated. There are two kinds of messages exchanged between data items the responsibility $r(i, k)$ and the availability massage $a(i, k)$, that reflects the accumulated evidence for how well-suited item x_k is to serve as the exemplar for item x_i, and reflects the accumulated evidence for how appropriate it would be for item x_i to choose item x_k as its exemplar.

$$r(i,k) = S(x_i, x_k) - \max_{j \neq k}\{S(x_i, x_j) + a(i,j)\} \qquad (1)$$

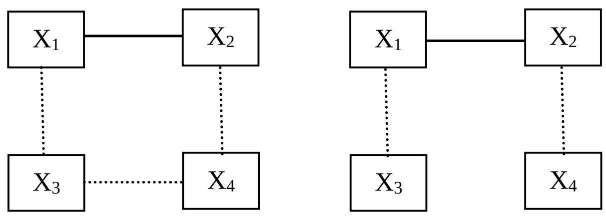
$$a(i,k) = \begin{cases} \sum_{i' \neq k} \max[0, r(i',k)] & i = k \\ \min\left[0, r(k,k) + \sum_{i' \notin \{i,k\}} \max[0, r(i',k)]\right] & i \neq k \end{cases} \qquad (2)$$

where, $S(x_i, x_j)$ denote the similarity between the data items x_i and x_j , with $i \neq j$.

Affinity propagation is an unsupervised clustering method. Semi-supervised clustering algorithms make use of the partially labeled data by using a limited number of constraints. The issue of selecting pairwise queries wisely to provide a precise clustering assignment is covered in this section. Using the neighborhood concept—where neighborhoods include labeled examples of various clusters depending on pairwise constraints—we choose the active constraints. By picking the most illuminating examples and investigating their connections to the communities, we broaden the neighborhoods. We summarize our strategy in Algorithm 1.

In order to create C clusters from a set of data points $X=\{x_1,..., x_n\}$, we can find a set of m neighborhoods $N = \{N_1, \cdot \cdot \cdot ,N_m\}$, where $m \leq C$. Imagine the data represented as a graph, with edges denoting must-link restrictions and vertices representing data instances. The neighborhoods are just the connected parts of the graph with cannot-link constraints between them. They are represented by the notation $N_i \subset X$, $i \in \{1, \cdot \cdot \cdot , m\}$.

Two examples that clarify how the neighborhoods can be formed from a set of pairwise constraints are shown in Fig. 1. Data instances are represented by nodes, must-link constraints are shown by solid lines, and cannot-link constraints are shown by dashed lines. Take note that there must be a cannot-link constraint between every neighborhood and every other neighborhood. Therefore, Fig. 1(b) only has two known neighborhoods, which might be either $\{x_1, x_2\}$, $\{x_3\}$ or $\{x_1, x_2\}$, $\{x_4\}$, but Fig. 1(a) has three neighborhoods: $\{x_1, x_2\}$, $\{x_3\}$, and $\{x_4\}$.



(a) (b)

Fig. 1. Two examples of neighborhoods based on pairwise constraints.

To determine the most informative points, let's consider a labeled dataset L consisting of pairs $\{(x_1, y_1), (x_2, y_2), . . . , (x_l, y_l)\}$, where y_l represents the cluster label of the data item x_l, along with an unlabeled set U containing data items $x_{l+1}, x_{l+2}, . . . , x_{l+u}$. Let E denote the set of exemplars in the dataset. Given a labeled sample x_i $(1 \leq i \leq l)$ and an unlabeled data item x_j $(l + 1 \leq j \leq u)$, we can identify two scenarios where the labeled sample might be associated with the unlabeled data item following execution of the AP algorithm:

1) If the unlabeled data item x_j adopts the labeled sample x_i as its cluster exemplar, and the message $a(x_i, x_i) + r(x_i, x_i)$ is positive (indicating $x_i \in E$), and x_i is the $\max\{a(x_j, x_k) + r(x_j, x_k)\}$ for each $k = \{1, 2, . . .,n\}$.

2) If the labeled sample x_i selects the unlabeled data item x_j as its cluster exemplar, and the message $a(x_i, x_i) + r(x_i, x_i)$ is negative (indicating $x_i \notin E$), and x_j is the $\max\{a(x_i, x_k) + r(x_i, x_k)\}$ for each $k = \{1, 2, . . .,n\}$.

If either of these conditions is met, the unlabeled data item x_j is deemed most similar to the labeled sample x_i. Consequently, x_j is chosen and assigned the label of x_i, effectively selecting the most similar unlabeled data item x_j as follows:

$$x^* = \begin{cases} x_j & if\ x_i = \max_{1 \leq k \leq n}\{a(x_j,x_k) + r(x_j,x_k)\}\ and\ x_i \in E \\ & if\ x_j = \max_{1 \leq k \leq n}\{a(x_i,x_k) + r(x_i,x_k)\}\ and\ x_i \notin E \end{cases} \quad (3)$$

where, x^* is the selection of the unlabeled point from the set U follows the operational principles of the AP algorithm

Once the most informative point is chosen, it is queried against the existing neighborhoods to ascertain its membership. To optimize query efficiency, the initial query is directed towards the neighborhood with the highest likelihood of containing x^*. This approach minimizes the overall number of queries required. The determination of a point x^* likelihood of belonging to a specific neighborhood N_i is based solely on its interactions with labeled points. This likelihood is estimated by averaging the similarities between x and the instances within N_i, as expressed by the formula:

$$p(x \in N_i) = \frac{\frac{1}{|N_i|} \sum_{x_j \in N_i} S(x,x_j)}{\sum_{p=1}^{l} \frac{1}{|N_p|} \sum_{x_j \in N_p} S(x,x_j)} \quad (4)$$

Here $S(x, x_j)$ represents the degree of similarity between point x and point x_j, $|N_i|$ denotes the amount of points in the neighborhood N_i, and l signifies the overall amount of neighborhoods.

Line 4 traverses the neighborhoods in decreasing order based on $p(x^* \in Ni)$, $i \in \{1, \cdot \cdot \cdot, l\}$, or the likelihood of x^* belonging to each neighborhood. We can find the neighborhood of x^* with the fewest number of searches by using this query order. With just one query, we can end if a must-link is returned. If not, the following question should be directed towards the neighborhood with the next highest likelihood of having x^*. A new neighborhood will be generated using x^* (lines 4–15) once this procedure is repeated until a must-link constraint is returned or we have a cannot-link constraint against all neighborhoods.

Algorithm 1. Affinity Propagation with Active Constraints

1. Start with a single neighborhood N_1 containing a randomly chosen instance x and set the number of queries q to 0.
2. While $q<Q$
3. Select the most informative point x^* to query using Equation 3;
4. **For** $N_i \in N$ ordered by decreasing probability of x^* belonging to Ni;
5. Query x^* against any data point x_i belonging to N_i;
6. q++;
7. Update the set of constraints according to the results of the queries;
8. **If** $ML(x^*, x_i)$ exist
9. $N_i = N_i \cup x^*$
10. Break;
11. else
12. make a new neighborhood containing the point x^*;
13. End if
14. End while

IV. Experimental Results

This section presents the datasets, evaluation metrics, Constraint selection strategies, and outcomes of the study. The effectiveness of the proposed method is explained through comparisons with several state-of-the-art algorithms across different scenarios to highlight its superiority.

A. Datasets

In this section, the datasets utilized are presented. For our experiments, real datasets were utilized. Hence, these datasets are labelled with instances, attributes, and numbers of clusters as described in Table I.

TABLE I. THE DATA SETS USED IN THE EXPERIMENTS

Dataset	# Instances	#Attributes	#Clusters
Glass	214	10	6
Ecoli	336	8	8
Ionosphere	351	34	2
Liver	345	6	2
Breast	683	9	2
Yeast	1484	8	10
Segment	2310	19	7
Magic	19020	10	2

B. Evaluation Metrics

We used pairwise F-measure and Normalized Mutual Information (NMI) as clustering validation metrics to evaluate the effectiveness of the approaches. NMI evaluation metric takes into consideration the clustering assignment and class label as random variable. Hence, the metric measures the common information between dual random variables. Therefore, this information will be normalized to zero-to-one range by the metric. *NMI* is computed as follows:

$$NMI = \frac{I(X;Y)}{\big(H(X)+H(Y)\big)/2}$$

In this context, $H(Y)$ represents Shannon entropy of Y, $H(Y|X)$ denotes conditional entropy of Y given X, and $I(X; Y)$ signifies the mutual information shared between the variables X and Y.

Pairwise F-measure was assessed in order to gauge clustering performance even more. Recall and precision were the sources of this statistic [16]. By comparing the predicted pairwise relationship between instance pairs to the ground truth class labels relationship comparison, the measure assesses one's predictive ability. Therefore, the harmonic mean of precision and recall is the common definition of F-measure. Thus, after obtaining a clustering result, we calculate the F-measure in the manner mentioned below:

$$Precision = \frac{n_c}{n_s}$$

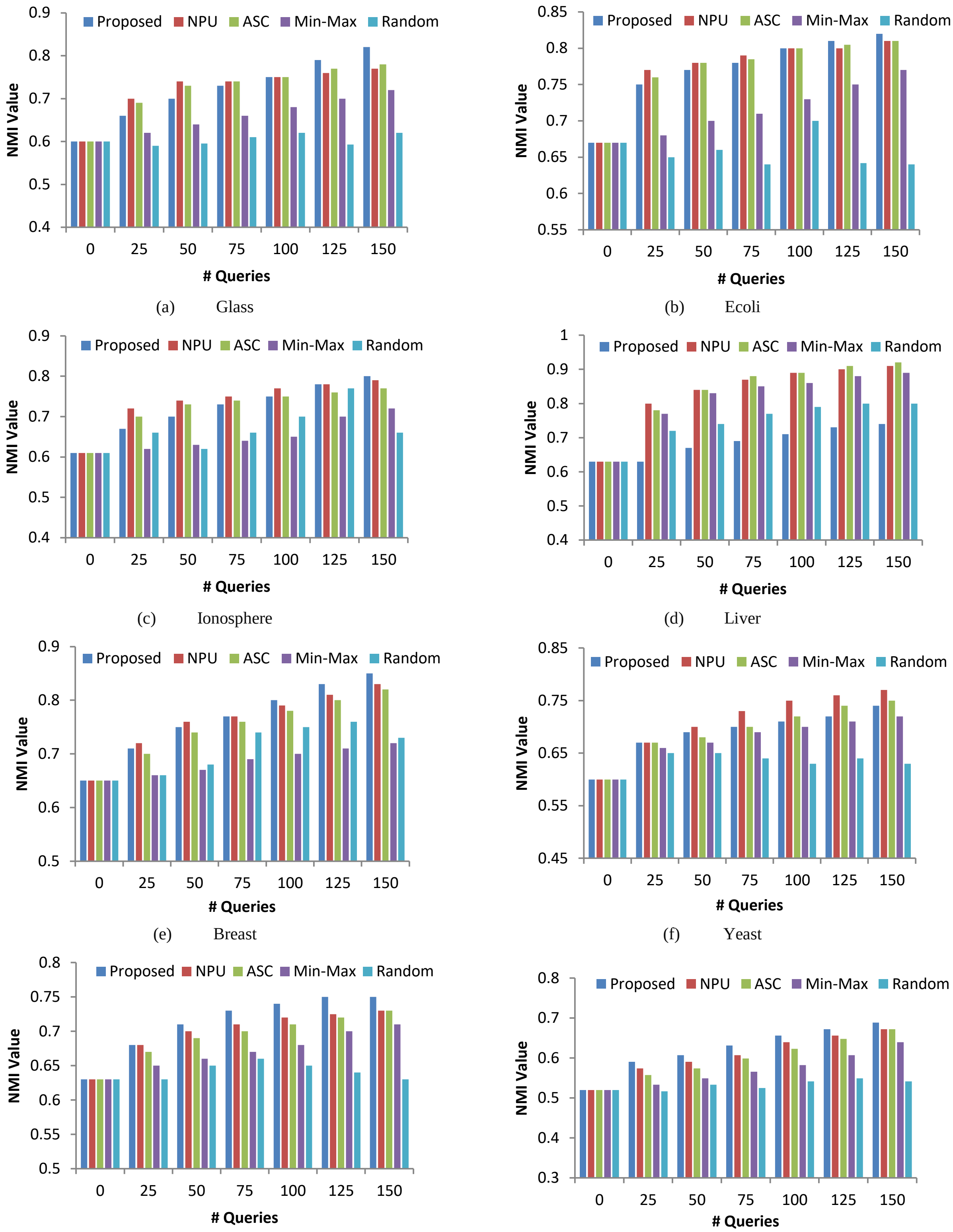
$$Recall = \frac{n_c}{n_f}$$

$$F - measure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

C. Constraint Selection Strategies

In all experiments, the following strategies are considered for selecting constraints:

- Random: This strategy entails a completely arbitrary selection of constraints. It involves generating a set of Must-Link (ML) and Cannot-Link (CL) constraints by comparing the labels of randomly chosen objects.
- Min-Max: This method follows a neighborhood-based approach and operates in two phases [18]. First, it creates a set of disjoint neighborhoods, Then, it incrementally expands these neighborhoods using a distance-based criterion.
- ASC: this method utilizing the neighborhood graph and formulating queries based on the constraint utility function. ASC relies on a pair of parameters, namely the threshold (θ) and the number of nearest neighbors (k). In accordance with their method, these parameters are set to $\lfloor (k/2) + 1 \rfloor$ and 6 respectively [8].
- NPU: This method is grounded in the uncertainty-based principle, employing a neighborhood-based strategy [16].



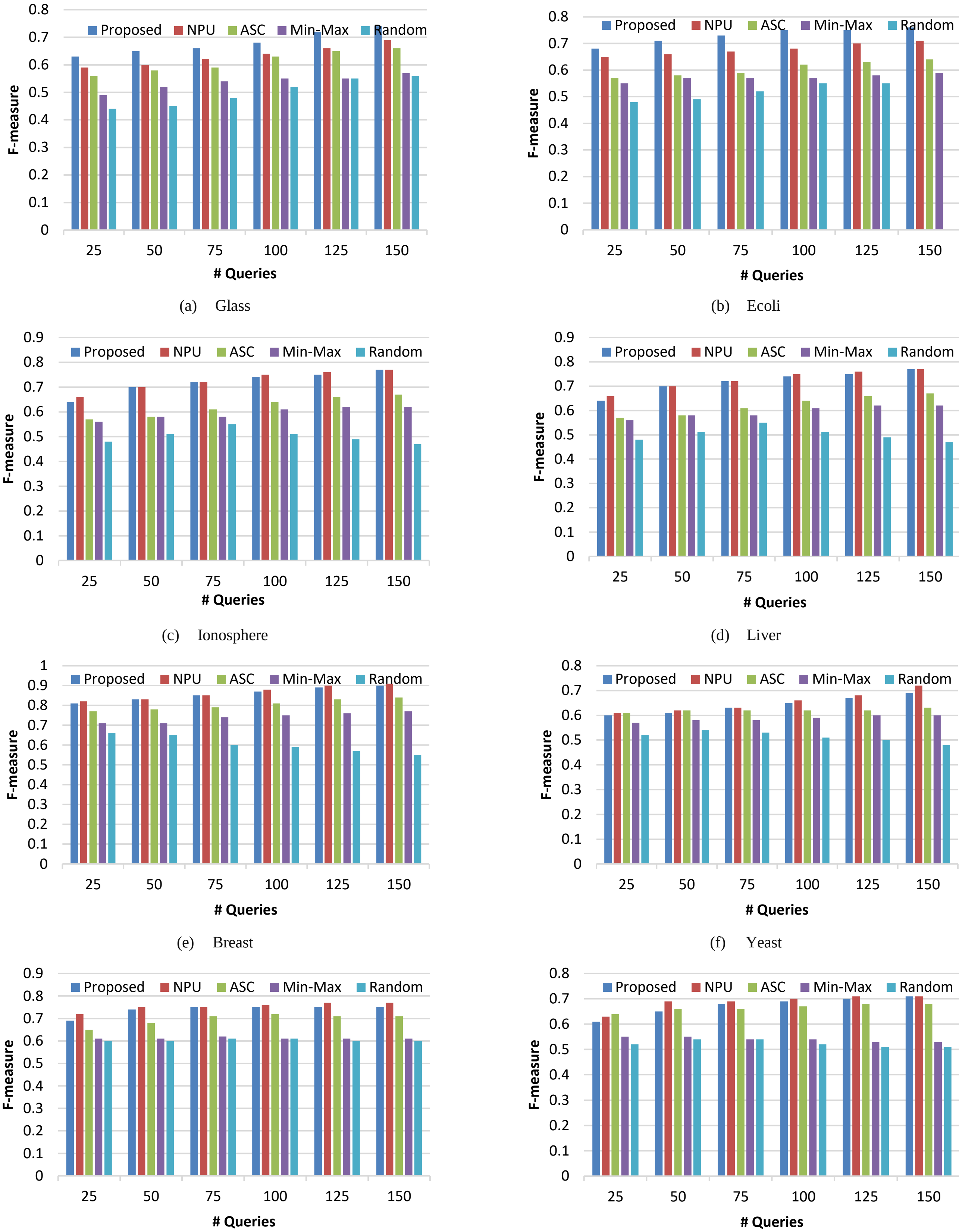
(a) Glass (b) Ecoli

(c) Ionosphere (d) Liver

(e) Breast (f) Yeast

(g) Segment (h) Magic

Fig. 2. Comparison of the suggested algorithm's clustering outcomes in NMI with various constraint selection techniques.



(a) Glass

(b) Ecoli

(c) Ionosphere

(d) Liver

(e) Breast

(f) Yeast

(g) Segment

(h) Magic

Fig. 3. Comparison of the suggested algorithm with other constraint selection techniques based on pairwise F-measure.

D. Performance Analysis Based on NMI

For the performance evaluation in this section, four algorithms were utilized for cross-comparison with our proposed algorithm. These algorithms are Random, min-max, ASC, and NPU. All these algorithms are active learning algorithms. Hence, the four datasets which are Glass, Ecoli, Segment, and Magic are used for the experiment. From Fig. 2, with 150 constraints on Glass dataset, the proposed algorithm performed better with 0.8 NMI, followed by ASC, NPU, Min-Max, and Random, respectively. The result is quit the same on Ecoli, Segment, and Magic datasets. One of the things worth noting is that the proposed algorithm is consistently effective. Meaning, it persistently outperformed all the algorithms compared with in all experiments. This conclusion is driven based on our general observation of Fig. 2.

In general, with respect to NMI evaluation, the proposed algorithm is more effective by large. It is important to also note that Random is the least effective algorithm. This is particularly due to the random selection of constraints by the algorithm in contrast to the other algorithms.

E. Performance Analysis Based on F-measure

The result in this section is given based on our evaluation using F-measure. Hence, the result is given of the comparison with other methods with respect to the datasets in Fig. 3. Hence, from Fig. 3, with focus on Glass datasets, the reader can see that the proposed algorithm surpasses the compared algorithms on all constraints. With respect to Ecoli dataset, our proposed algorithm also outperformed the compared algorithms with great margin. We observed that the proposed algorithm is does not have a good performance in large dataset like Magic with small number of constraints. However, the proposed algorithm achieves better performance result when we have a large number of constraints.

However, on Segment dataset, our proposed algorithm was outperformed by NPU algorithm 25, 125, and 150 constraints. Hence, looking at the result carefully, on 50, 75, and 100 constraints, the proposed algorithm outperformed all the compared algorithms. Furthermore, on Magic dataset, NPU and ASC outperformed the proposed algorithm on 25 and 50 constraints. However, from 75 constraints and above, the proposed algorithm outperformed all the compared algorithms as presented in Fig. 3.

V. Conclusion and Future Work

In this study, we introduced an approach to improve the semi-supervised clustering algorithms that select the active pairwise constrained with affinity propagation clustering algorithm. Initially, the most informative points are generated using the AP algorithm, and subsequently, only the points are chosen to compose the neighborhoods and generate the final clustering outcomes. Additionally, in acquiring pairwise constraints, we replaced random selection with an active learning strategy, resulting in more representative constraints. Our algorithm was applied to eight datasets from UCI datasets, with the performance evaluated using NMI and F-measure metrics. The experimental findings demonstrate the superiority of our proposed method over other clustering algorithms. In future work, we would like to work on the problem of incremental growing constraint set for streaming data. To address this problem, we are interested to apply an incremental semi-supervised clustering method.

Acknowledgment

This work was funded by the University of Jeddah, Saudi Arabia, under grant No. (UJ-20-094-DR). The authors, therefore, acknowledge with thanks the university's technical and financial support.

References

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

[2] Davidson, I and Ravi, S. S. "Clustering with constraints: feasibility issues and the k-means algorithm", In *proceedings of the 5th SDM*, 2005. pp. 138–149.

[3] Atwa, W. and Almazroi, A.A. Active Selection Constraints for Semi-supervised Clustering Algorithms. Int. J. Inf. Technol. Comput. Sci., 2020, pp 23–30.

[4] Davidson, I., Wagstaff, K. and Basu, S. "Measuring constraint-set utility for partitional clustering algorithms," *Knowledge Discovery in Databases*, 2006, pp. 115–126.

[5] Greene, D. and Cunningham, P. "Constraint selection by committee: An ensemble approach to identifying informative constraints for semi-supervised clustering," *European Conference on Machine Learning*, 2007, pp. 140–151.

[6] Chen, D.W. and Jin, Y.H. An active learning algorithm based on Shannon entropy for constraint-based clustering. IEEE Access, 8, 2020, pp.171447-171456.

[7] S. Baohua, J. Juan, Q. Feng, L. Daoguo, Y. Yanming, A. Gholamreza, Semi-supervised hierarchical ensemble clustering based on an innovative distance metric and constraint information, Engineering Applications of Artificial Intelligence, 2023, vol.124, pp.106571.

[8] Huang, R. and Lam, W. "Semi-supervised Document Clustering via Active Learning with Pairwise Constraints," in *International Conference on Date Mining*, 2007, pp. 517–522.

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

[10] B. Frey, and D. Dueck, 'Clustering by passing messages between data points', Science, 2007, pp. 972–976.

[11] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in Icml, 2001, vol. 1, pp. 577–584.

[12] S. S. Rangapuram and M. Hein, "Constrained 1-Spectral Clustering.," in AISTATS, 2012, vol. 30, p. 90.

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

[14] N. Shental, A. Bar-Hillel, T. Hertz, and D. Weinshall, "Computing Gaussian mixture models with EM using equivalence constraints," in *Advances in neural information processing systems*, 2004, pp. 465–472.

[15] M. Bilenko, S. Basu, and R. J. Mooney, "Integrating constraints and metric learning in semi-supervised clustering," in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 11.

[16] S. Xiong, J. Azimi, and X. Z. Fern, "Active learning of constraints for semi-supervised clustering," *IEEE Trans. Knowl. Data Eng.*, 2013, vol. 26, no. 1, pp. 43–54.

[17] M. C. Fernandes, T. F. Covões, and A. L. V. Pereira, "Improving evolutionary constrained clustering using Active Learning," *Knowledge-Based Syst.*, 2020, vol. 209, p. 106452.

[18] P. K. Mallapragada, R. Jin, and A. K. Jain, "Active query selection for semi-supervised clustering," in *2008 19th International Conference on Pattern Recognition*, 2008, pp. 1–4.

[19] H. T. Nguyen and A. Smeulders, “Active learning using pre-clustering,” in *Proceedings of the twenty-first international conference on Machine learning*, 2004, p. 79.

[20] V.-V. Vu, N. Labroche, and B. Bouchon-Meunier, “Active learning for semi-supervised k-means clustering,” in *2010 22nd IEEE International Conference on Tools with Artificial Intelligence*, 2010, vol. 1, pp. 12–15.

[21] W. Zhao, Q. He, H. Ma, and Z. Shi, “Effective semi-supervised document clustering via active learning with instance-level constraints,” *Knowl. Inf. Syst.*, 2012, vol. 30, no. 3, pp. 569–587.

[22] X. Wang and I. Davidson, “Active spectral clustering,” in *2010 IEEE International Conference on Data Mining*, 2010, pp. 561–568.

[23] Y. Li, Y. Wang, D.-J. Yu, N. Ye, P. Hu, and R. Zhao, “Ascent: Active supervision for semi-supervised learning,” *IEEE Trans. Knowl. Data Eng.*, 2019, vol. 32, no. 5, pp. 868–882.

[24] Q. Lei and T. Li, "Semi-Supervised Selective Affinity Propagation Ensemble Clustering With Active Constraints," in *IEEE Access*, 2020, vol. 8, pp. 46255-46266.

[25] R. Hazratgholizadeh, M. A. Balafar, M. R. F. Derakhshi, "Active constrained deep embedded clustering with dual source", Applied Intelligence, 2022.

[26] L. Aronsson and M. H. Chehreghani, "Correlation Clustering with Active Learning of Pairwise Similarities". arXiv preprint arXiv. 2023, 2302.10295.