Research Article

Active Learning for Constrained Document Clustering with Uncertainty Region

M. A. Balafar ¹, **R. Hazratgholizadeh**,¹ and **M. R. F. Derakhshi**²

¹Department of IT, Faculty of Engineering, University of Tabriz, Tabriz, Iran

²Department of Computer, Faculty of Engineering, University of Tabriz, Tabriz, Iran

Correspondence should be addressed to M. A. Balafar; balafarila@yahoo.com

Received 9 January 2020; Accepted 1 May 2020; Published 20 May 2020

Academic Editor: Mohammad Hassan Khooban

Copyright © 2020 M. A. Balafar et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Constrained clustering is intended to improve accuracy and personalization based on the constraints expressed by an Oracle. In this paper, a new constrained clustering algorithm is proposed and some of the informative data pairs are selected during an iterative process. Then, they are presented to the Oracle and their relation is answered with “Must-link (ML) or Cannot-link (CL).” In each iteration, first, the support vector machine (SVM) is utilized based on the label produced by the current clustering. According to the distance of each document from the hyperplane, the distance matrix is created. Also, based on cosine similarity of word2vector of each document, the similarity matrix is created. Two types of probability (similarity and degree of similarity) are calculated and they are smoothed for belonging to neighborhoods. Neighborhoods form the samples that are labeled by Oracle, to be in the same cluster. Finally, at the end of each iteration, the data with a greater level of uncertainty (in term of probability) is selected for questioning the oracle. In order to evaluate, the proposed method is compared with famous state-of-the-art methods based on two criteria and over a standard dataset. The result demonstrates an increased accuracy and stability of the obtained result with fewer questions.

1. Introduction

Clustering is one of the main important methods in the background of machine learning [1] and can be applied to different datasets such as the document set. In the common methods of clustering, there is no previous information, and as such, it is called the unsupervised learning method [2, 3]; however, in the real world, some information [4] is normally available, or we can obtain from Oracle. This information can be in different forms and can be used in the process of clustering [5–13].

If the information is presented as pairwise constraints (where a document pair must be in the same cluster (ML), while a document pair should not be located in the same cluster (CL)), and these pairwise constraints are used in the process of clustering, this method will be called pairwise constrained clustering [6, 14, 15]. Pairwise constraints can be useful in the clustering process in two ways: when enough informative pairwise constraints exist, where the accuracy

and efficiency of the clustering can be improved, and when we want to change the process of clustering and personalize it [10, 12, 16].

Generally, it is important to select a valuable data pair as an informative pairwise constraint. Active learning selects the informative pairwise constraints and sends them to Oracle for responding (Must-link/Cannot-Link) [15]. Active learning offers the greatest improvement and accuracy in clustering by saving time and cost for the minimum number of pairwise constraints [1, 11, 14, 17–20]. Active learning is performed frequently in the classification and has provided better results. However, active learning in clustering is used with limitation. The aim of these methods is to select a data pair not correctly clustered by the current clustering most of the time [14, 21–23].

Often, pairwise constraints are selected once, and they are sent to Oracle, while the current clustering cannot have a more effective role in the selection. For this reason, iterative methods and concepts such as neighborhoods and

uncertainty are used for the selection of informative pairwise constraints [14, 24–26].

The main objective of this study is to present active learning for pairwise constrained document clustering with uncertainty region. Traditionally, active learning is used to select valuable data and fewer questions from the Oracle. Active learning depends on a variety of methods for better results. These methods are usually statistical and complex. There are less effective methods with simplicity and intuition. In this study, uncertainty region is an effective, simple, and novel method. Active learning has the most impact on the selection of valuable documents. In constrained document clustering, obtaining pairwise constraint requires a user with excessive amount of work to read the documents in question and indicate their relationship, which is feasible but time consuming. For this reason, we use the document dataset for evaluation. This dataset highlights the impact of active learning.

Since dataset is document clustering, preprocessing is the necessary step for achieving the best result. Document representation and word embedding constitute the core of this step. A common approach to represent the document is bag-of-words (N-gram by TF-IDF). This method is useful to achieve word frequency. However, structural and semantic information is ignored. Some methods, such as graph representation [27], wikifying, word net, and others are useful for enrichment and semantic representation. Nowadays, neural networks based on language methods significantly outperform traditional methods which can preserve the semantic relationship between the words of documents. For example, Word2vec is a simple and effective method for word (and document) representation and dimension reduction. This paper uses the two mentioned methods to improve constrained clustering results [3, 5, 11, 19, 28–33].

In this study, some data pairs as informative pairwise constraints are selected in each iteration and, based on a response received from Oracle the set of constraints, is updated. Iteration continues until the number of the queries from Oracle reaches zero. In each iteration, constrained clustering is performed. Then, SVM is used based on the label assigned to each data from the currently constrained clustering. Following the SVM model, the distance matrix and similarity matrix are created, respectively, based on the distance of each document from the hyperplane (HP) and cosine similarity between each semantic representation of documents such as word2vector [31, 32, 34]. Afterwards, the probability of each data belonging to every neighborhood is calculated based on these matrices. The similarity and degree

of similarity methods are implemented for obtaining the probabilities. Our method suggests a new concept such as the uncertainty region for expressing the degree of similarity. In each iteration, the boundary of this region is obtained.

Uncertainty estimation such as silhouette index is an approach in active learning which has been widely used, where the data point is ranked based on the level of uncertainty regarding their probability of belonging to neighborhoods. To select valuable data as an informative pairwise constraint, we use the silhouette index with smaller values representing a greater level of uncertainty [26]. Experimental results reveal the stability and improvement of accuracy over five different datasets and four famous state-of-the-art methods. Ultimately, our main contributions are summarized as follows:

- (1) Presentation of a new concept (the uncertainty region) to detect uncertainty instead of complicated methods in the literature
- (2) Creation of boundary of uncertainty region and penalty of violation from constraints automatically (by histogram thresholding)
- (3) Development of an adaptive and consensus ensemble method (local search for selecting uncertainty data pair items instead of global search) to balance clusters and achieve stable results

Concerning the organization of the paper, we address the related works in Section 2. In Section 3, the proposed method and “Materials and Methods” are expressed. Section 4 presents the experimental method, and Section 5 presents experimental results and discussion. In Section 6, we conclude the paper and present future works.

2. Related Works

One of the methods which has been rarely considered is the use of a support vector machine, deep learning concept, and neighborhoods, especially in constrained clustering [22, 31, 35]. First, we explain these concepts and then introduce methods similar to our proposed method.

Generally, if $\{x_i, x_j, x_k\}$ are three data points of a dataset, Oracle response is as {Must-Link, Cannot-Link} and current clustering label is expressed with lab . Then, equation (1) must hold true in order to have problem-free neighborhood concepts [14, 24, 26]:

$$\begin{aligned}
 & * (x_i, x_j, \text{Must} - \text{Link}) \wedge (x_i, x_k, \text{Must-Link}) \implies (x_j, x_k, \text{Must} - \text{Link}), \\
 & * (x_i, x_j, \text{Cannot} - \text{Link}) \wedge (x_i, x_k, \text{Cannot} - \text{Link}) \implies (x_j, x_k, \text{Cannot} - \text{Link}), \\
 & * (x_i, x_j, \text{Must} - \text{Link}) \iff lab_i = lab_j, \text{ in same cluster,} \\
 & * (x_i, x_j, \text{Cannot} - \text{Link}) \iff lab_i \neq lab_j, \text{ not in same cluster.}
 \end{aligned} \tag{1}$$

Each neighborhood includes data samples, which are in the same cluster. The essential assumption in the neighborhood is that the existing data in different neighborhoods must lie in different clusters. Neighborhoods can be shown as $\text{neighbor} = \{N_1, N_2, N_3, \dots, N_h\}$, where a number of h neighborhoods exist [24, 26].

Neighborhoods are used as they are economical in questioning the Oracle such that, after selecting the informative data in each iteration, we can present those data only versus one of the data samples of each neighborhood to Oracle. If the result is Must-Link, we will add this data to that neighborhood, and it will be Must-link with all the data available in that neighborhood. Now, if we do not have Must-link versus any of the members of neighborhoods, then a new neighborhood must be created and those data must lie in a new neighborhood [26].

Deep learning can train a huge dataset and outperform the traditional methods. This method is useful in dimension reduction and preserving the semantics as well as the structure of the dataset. There is little research which has applied deep learning methods in constrained clustering. Via this method, we can apply the dimension reduction and calculate the similarity matrix and others. Our paper used the deep learning method effectively in the process of a constrained method [11, 13, 31, 32].

Here, the dataset with $\mathcal{D} = \{x_1, x_2, x_3, \dots, x_n\}$ has the constraints with data pairs with Must-link and Cannot-Link relationship. In this case, the objective function can be changed for combination and applying pairwise constraints in the clustering process. In this change, a penalty can be considered for the violation from pairwise constraints [17]. For example, PCKmeans that the algorithm [36] has been able to use both standard objective function and a penalty for the violation of constraints, by changing the objective function. These two sections together constitute the objective function and are locally minimized.

On the contrary, in COPKmeans algorithm, no violation occurs from pairwise constraints. This method is called hard clustering, which is contrary to the previous clustering algorithm called soft clustering algorithm [36].

Along with the algorithms that address the development of new methods of constrained clustering, there are weaker methods in order to select the informative pairwise constraint [6–8, 36]. Active learning is widely used for classification issues when the label-having data is few for the training set [9, 11, 13, 18, 19, 22, 35, 37].

In this regard, the first active learning algorithm for constrained clustering was first developed by Basu et al. [36]. This algorithm has two main phases of exploration and consolidation. Gradually, in the first phase, the data are selected based on the farthest-first strategy. After selection, the data are presented to Oracle with a data sample of each neighborhood; if it is not located in any neighborhood, then a new neighborhood will be created. The second phase selects data iteratively and randomly, after which the data are presented to the Oracle with a data sample of each neighborhood until it lies in one neighborhood. In the first phase, the objective is to develop the number of neighborhoods, while in the second phase, the objective is to develop the

number of data samples in each neighborhood. Obviously, this method is the basis of other methods and there have been some developments of this algorithm. The examples include [17], where informative data selection was not performed randomly.

Greene and Cunningham [24] performed informative data selection using another method. In this algorithm, two main phases, similar to the previous algorithm, have been used. First, the dataset is clustered with different algorithms; then, for each data pair, according to the frequencies it lies in the same cluster (at all times), the similarity matrix is constituted. At this stage, two thresholds are determined from the values of this matrix, and data pairs are divided into three categories. The section with values higher than thresholds is selected as data pairs of Must-link whose transitive closure forms the neighborhoods. In the first phase, the mean of values is calculated for each neighborhood in terms of the existing data; then, initial clustering is formed based on the correspondence of data to these means. In the second phase, from the similarity matrix, for each data, the probability of belonging to each cluster is calculated and the data with the greatest uncertainty for belonging to clusters are selected as informative data.

Xiong et al. [26] introduced the most appropriate and close framework to our proposed method in terms of active learning background. In this paper, constrained clustering is used as a black box, and in each iteration, the results of only the performed clustering are applied. In each iteration, the result of clustering is determined as a class label for each data; then, the method of random forest classification is used. The ratio of the number of times every data pair is placed in one leaf to the total number of model iteration is regarded as an element of the similarity matrix between every data, where the similarity matrix is obtained. The probability for each data belonging to neighborhoods is calculated by the similarity matrix. Then, informative data for each question are obtained based on the probability obtained for each data and neighborhoods from uncertainty sampling estimation methods such as entropy and cost expectation mean.

Recently, Xiong et al. [38] developed a new online framework for active clustering with model-based uncertainty detection. This method uses semisupervised spectral clustering as a black box, which selects pairwise constraints as a clustering process, based on the uncertainty detection principle. The main idea in this paper is based on concepts of “certain sample sets” and “estimate the uncertainty.” Certain sample sets are approximately similar to neighborhoods’ sets. For estimating the uncertainty, a novel method approximates the first-order model, which decomposes the expected uncertainty into two components: a gradient and a step-scale factor. Calculations of this framework are complicated and time consuming in terms of runtime. This framework is hard for preserving semantic and dimension reduction in unstructured datasets such as a document.

Oliveira et al. [39] proposed new hybrid methods that used random key genetic algorithm with local search heuristic and column generation with path relinking. They found that genetic algorithm with local search can act as an

alternative and efficient method to solve the constrained clustering problem. Yang et al. [40] introduced a theoretical effect of the diversity and quality of the ensemble and then proposed a unified framework to solve the clustering ensemble selection problem with three criteria metrics. Wei et al. [41] introduced a semisupervised clustering ensemble approach which involved both pairwise constraints and metric learning. In this method, via supervised information, the method generated different base clustering partitions using constraint-based semisupervised clustering and metric-based semisupervised clustering, respectively. Then, consensus function smoothed the result of each independent clustering. Yu et al. [42] developed a new ensemble clustering with active learning and selected constraint projection. In this method, first, a random subspace dataset was provided, after which with constraint set, high-dimensional data were mapped to low-dimensional space. After provision of subspace and dimension reduction, different weights were generated for each constrained cluster. Finally, with consensus function, each result of clustering was ensembled.

Another category of algorithms also exists in this field including production of pairwise constraints actively and iteratively [8, 18, 19, 24, 43], genetic heuristic-based algorithms [39], communication between constraints for enriching constraint sets [44], and constraint space transfer with kernel [45]. These methods have often used previously published algorithms as a black box trying to reduce uncertainty with novel methods. The methods, such as uncertainty sampling-based [24, 26], commission (or ensemble) as well as hybrid method [40, 46], and lowering error rate in the main model along with distance from hyperplane in SVM [30], are used in active learning and constrained algorithms [15, 19]. These methods have been used in the literature in different ways [8, 26].

The main drawbacks of the mentioned algorithms included the following: unstable results, ignoring semantic representation, using a weak method to measure similarity, heavy calculations, using random selection broadly, weak dimension reduction, and using weak uncertainty detection methods.

3. Materials and Methods

Since clustering dataset is a document, first, it is necessary to convert the document set to the document-term matrix. For this purpose, preprocess should be used such as removing the empty document, numbers, and nonstop terms. In order to extract all terms of a document, the token process is required, in which a document is tokenized into a batch of terms, with each term in documents being given a weight. In the following, for decreasing the dimensions of this matrix, only informative terms are preserved.

Furthermore, word2vector methods are used for preserving the semantics and structure of the document dataset. Word2vec method uses a raw dataset to generate a vector for each word in the document. Then, simply by an averaging vector of each word in the document, the document vector is generated. This method is used to create a similarity-document-term matrix. A row of this matrix consists of the

document vector where the size of the dimension is equal to the size of the document-term matrix [32].

Afterwards, we use PCKmeans clustering as a black box. In this paper, two essential changes have been applied to PCKmeans: (i) in the section of initialization and (ii) in the section of calculating centers of clusters. Furthermore, the penalty of violation from constraints is created automatically [12, 47].

3.1. Problem Preliminaries. The set of documents is shown as $\mathcal{D} = \{x_1, x_2, x_3, \dots, x_n\}$, in which x represents the document and $i = 1, \dots, n$. Then, by applying preprocessing, the terms of these documents are converted into the weights with different values by TF-IDF (which has the best result in this case) [12, 14, 19, 29, 33, 48]. In this case, each document can be shown as $x = \{w_1, w_2, w_3, \dots, w_t\}$, $b = 1, \dots, t$. In this set, w represents the weight of the intended term, which is obtained as follows:

$$w_{ib} = \text{tf}_{ib} \times \text{idf}_{ib} = \text{tf}_{ib} \times \log_2 \frac{n}{\text{df}_{ib}}. \quad (2)$$

In this formula, tf is the frequency of the term b in document i and df refers to the frequency of the number of documents, in which this term exists. In order to reduce the term dimensions of the matrix, mean-tfidf method is used; initially, the weight mean is calculated for each term. Then, the terms that are higher than the weight mean remain in the matrix while other terms are removed. Equation (3) shows this method:

$$\text{Mean}_{\text{tfidf}}(\text{term}_b) = \frac{1}{n} \sum_{i=1}^n \text{tfidf}_{ib}, \quad (3)$$

After creating the distance matrix and reducing its dimension (in addition to using word2vector to create similarity matrix), Murkowski distance method has been used in the clustering algorithm in this paper. This method is one of the most famous methods for clustering.

3.2. The Proposed Clustering Algorithm. Support vector machine based on pairwise constrained clustering algorithm—SVBPCKmeans—is represented in Algorithm 1. The objective function of this algorithm is minimized locally, as with the PCKmeans method. In the proposed method, the penalty for violating constraints is calculated and normalized at each stage (in contrast to PCKmeans).

In this algorithm, first, C as the set of pair constraints, neigh as the set of the neighborhood, and h as the current number of neighborhoods are initialized. Then, Algorithm 2, known as “Cons_set_initial,” is called only once at the beginning of the algorithm. In the following, the results obtained from Algorithm 2 are utilized. Afterwards, *while* section will continue until q (question from Oracle) does not reach zero.

The center of each neighborhood is calculated at the beginning of each loop, with these centers of neighborhoods introduced during the clustering algorithm process as the initial value of clustering centers. Note that if the number of

Input: Document-term matrix $\{M_{ib}\}_{i=1,\dots,n}^{b=1,\dots,t}$, Similarity-document-term matrix $\{SM_{ib}\}_{i=1,\dots,n}^{b=1,\dots,t}$, Dataset $\mathcal{D} = \{x_1, x_2, x_3, \dots, x_n\}$, Number of cluster K , Total number of query q , Label assigned from current pairwise constrained clustering lab, Current number of neighborhood h , Neighborhoods $\text{neigh} \{N_h\}_{h=1}^k$, Constraint set C , Must link ML , cannot link CL , Centroid of cluster p at t iteration $\mu_p^{(t)}$, penalty of must link W_m , penalty of cannot link W_c .

Output: K document cluster $\{Xp\}_{p=1}^K$

Initializations: $C = \emptyset$, $h = 0$, $\text{neigh} \{N_h\}_{h=1}^k = \emptyset$, $t = 0$

$[h, \text{neigh}, C, W_m, W_c] = \text{Cons_set_initial}(q, M, \text{neigh}, h, K)$

While true

Find centroid of $\text{neigh} \{N_h\}_{h=1}^k$ //average of each neighborhood

If $h < k$, then find $k-h$ random point, not in neighborhoods

Initial $\left\{ \mu_p^{(0)} \right\}_{p=1}^K$ with centroid of $\text{neigh} \{N_h\}_{h=1}^k$ and random chosen point

Repeat until convergence//not changed pairwise constrained clustering

A: assign_cluster:

Assign each data point $x_i = M_i$ to the cluster p^* such that:

$$p^* = \arg \min (1/2 \|x_i - \mu_p^{(t)}\|^2 + \sum_{(x_i, x_j) \in ML} W_m I[p \neq \text{lab}_j] + \sum_{(x_i, x_j) \in CL} W_c I[p = \text{lab}_j])$$

B: estimate_means:

$$\left\{ \mu_p^{(t+1)} \right\}_{p=1}^K = \alpha \left\{ 1/|X_p^{(t)}| \sum_{x \in X_p^{(t)}} x \right\}_{p=1}^K + (1 - \alpha) \text{centroid of } \text{neigh} \{N_h\}_{h=1}^k$$

C: $t \leftarrow t + 1$

If $q = 0$, then return.

$[h, \text{neigh}, C, W_m, W_c] = \text{Cons_set_develop}(q, M, h, \text{neigh}, K, \text{lab})$

Endwhile

ALGORITHM 1: SVBPCMeans.

Input: document-term matrix M , Similarity-document-term SM , Number of cluster K , Total number of query q , Current number of neighborhood h

Output: $h, C, \text{neigh} \{N_h\}_{h=1}^k, W_m, W_c$

Initializations: $W_m = W_c = 0.001$

$h = h+1$ and set random data point x to $\text{neigh} \{N_h\}$

While $q > 0$ and $h < K$

Apply k -means to M and get centroid $\left\{ \mu_p \right\}_{p=1}^K$

Ind_point = Find k nearest points to the centroid $\left\{ \mu_p \right\}_{p=1}^K$

For each Ind_point x

For each $\text{neigh} \{N_h\}_{h=1}^k$

Query x versus one member of $\text{neigh} \{N_h\}$ and $q = q - 1$

If answer is Must-link, then

A: Add x to the corresponding neigh

B: Update pair constraint C and break

Endfor

If x cannot _link answered to all neigh , then

A: $h = h+1$

B: Start new neigh and add x to $\text{neigh} \{N_h\}$

C: Update Pairwise Constraint C

Endfor

Endwhile

ALGORITHM 2: Cons_set_initial.

neighborhoods is lower than the required number of clustering centers, the other centers must be selected randomly.

In the next section, we enter an iterative process which will continue until the objective function with the violation penalty values does not reach convergence. In this iterative process, centers of new clusters will be obtained in Algorithm 1. New cluster centers will be smoothed with neighborhood centers obtained from the previous stage by a coefficient. The main

reason of this smoothing is to find the chance to establish a balance between the centers of newly created clusters and centers of the neighborhood at each stage. The process is terminated after convergence, where Algorithm 3, known as “Cons_set_develop,” is called. The results of this algorithm are the same as those of Algorithm 2, but they are used at each repetition. Figure 1 depicts an overview of the proposed algorithm to explain the main steps of the proposed algorithm.

FIGURE 1: The overview of the proposed algorithm.

3.3. Active Learning Method. The task of Algorithm 2 is to explore the neighborhoods. This algorithm continues until the question of Oracle is allowed or the number of neighborhoods is not larger than that of clusters. The first neighborhood with $h = 1$ will be formed using random data. The strategy in this algorithm is to use the centers of clusters obtained from a simple clustering algorithm, such as k-means. In this method, clustering is performed on the dataset; then, centers of clusters are obtained and finally the nearest data to the centers of clusters are selected.

Each of the nearest data will be presented to the Oracle in an iterative mode versus each data sample of neighborhoods. If the answer is ML, the nearest data will be appended to the corresponding neighborhood and update the constraint set and break; otherwise, a new neighborhood is formed if there are no neighborhoods with ML response.

In this strategy, the main objective of the algorithm is to find the maximum number of neighborhoods, benefiting from the clustering algorithm. For example, the centers of clusters have a Cannot-link relation. Thus, this method is better than the strategies such as selection of the farthest-first points or random points. In our method, fewer questions are required to reach the maximum number of neighborhoods, while the rest of the question can be used in the next algorithm, which is Algorithm 3.

Algorithm 3 tries to build each neighborhood with a balanced number. In this algorithm, the objective is to find informative data with Must-link response from Oracle. First,

Algorithm 4 is known as “informative-points.” This algorithm selects k informative data points equal to the number of clustering and send them to Algorithm 3. Thereafter, the distance between the informative data point and the centers of neighborhoods is calculated so that the distances are sorted in an ascending order and are sent to Oracle versus the data sample of each neighborhood. The goal is to find the Must-link with a minimum cost. Finally, the data with Must-link answer are added to the corresponding neighborhood, and all the sets are updated accordingly.

Note that Algorithm 4 can be mentioned as the main algorithm for selecting the informative data points. This algorithm also dynamically finds the penalty of violating Must-link under the name W_m and the penalty of violating Cannot-link called W_c . In this algorithm, we introduce a new concept known as the uncertainty region. This region is used for determining the degree of similarity which is the basis for documents belonging to neighborhoods. Indeed, the uncertainty region is a set of data pairs with greater uncertainty based on values in the distance matrix.

This algorithm assigns the label of the data obtaining from the current pairwise constrained clustering (in Algorithm 1) and considers them as the class. Then, it uses SVM classification for k times. In this way, the distance from the HP is calculated for the available data, and d_m matrix is calculated for each pair of data according to equation (5). The point for this method is that, in contrary to the common methods, the values of this matrix lie within the continuous

Input: document-term matrix M , Similarity-document-term SM, Number of cluster K , Total number of enquiry q , Label assigned from current pairwise constrained clustering lab, Number of neigh h
Output: C , neigh $\{N_h\}_{h=1}^k$, W_m , W_c
 $[W_m, W_c, \text{ind_point}] = \text{informative_points}(M, \text{SM}, \text{neigh}, C, h, K, \text{lab})$
For each point x in ind_point
Find distance between point x and each neighborhood centroid and sort neighborhoods distance ascending
For each neigh $\{N_h\}_{h=1}^k$
Query x versus one member of neigh $\{N_h\}$ and $q = q - 1$
If answer is Must-link, then
A: Add x to the corresponding neigh
B: Update C pairwise constraint
C: Break
Endfor
If $q = 0$ return
Endfor

ALGORITHM 3: Cons_set_develop.

Input: document-term matrix M , Similarity-document-term SM, Number of cluster K , Total number of enquiry q , Current number of neigh h , Label assigned from pairwise constrained clustering lab, Neighborhoods neigh $\{N_h\}_{h=1}^k$
Output: ind_point , W_m , W_c
Initializations: $M' = \{M_i, \text{lab}_i\}$
Run SVM classifier on M' , k times based on random chosen training set.
Compute normalized distance matrix d_M for every pair based on equation (5).
Compute normalized similarity matrix s_SM for every pair based on: equation (6).
Find m_d , m_c based on equations (7) and (8).
 $W_m = m_d$, $W_c = m_c$
For each $x_i = SM_i$ and neigh $\{N_h\}_{h=1}^k$, compute: $\text{sim_to_neigh}(x_i, N_h)$ based on equation (9).
For each $x_i = M_i$ and neigh $\{N_h\}_{h=1}^k$, compute: $\text{degree_toneigh}(x_i, N_h)$ based on equation (10).
For each $x_i = M_i$, compute $\text{uncert}(x_i)$ based on equation (14).
Sort clusters of current constrained clustering based on size in decreasing.
For each cluster respectively,
Find data point with the smallest value on $\text{uncert}(x_i)$ U .
If U is not in neigh $\{N_h\}_{h=1}^k$, then add it to ind_point .
If size of ind_point is k , then return.
Endfor

ALGORITHM 4: Informative_points.

interval $[0, 1]$ after normalization, which can create a high decision-making power. Next, we calculate the normalized matrix s_sm from the similarity-document-term matrix SM which is calculated for each pair of data according to equation (6). s_sm uses pretrained word2vector and cosine similarity method [18, 32]:

$$d_i = \frac{1}{k} \sum_{h=1}^k \frac{\text{HP}_{hp}(M_i)}{\|w_{hp}\|}, \quad (4)$$

$$d_M_{ij} = \begin{cases} \text{if label}(M_i = M_j) \text{ in SVM strategy: } |d_i - d_j|, \\ \text{if label}(M_i \neq M_j) \text{ in SVM strategy: } d_i + d_j, \\ \text{else } i = j: 0, \end{cases} \quad (5)$$

$$s_SM_{ij} = \begin{cases} \text{if } i \neq j \text{ or } (i, j) \in \text{CL: normalized similarity cosine} \\ (sm_i, sm_j), \\ \text{else } i = j \text{ or } (i, j) \in \text{ML: } 1. \end{cases} \quad (6)$$

In line 5 of Algorithm 4, the m_c and m_d are calculated from values on matrix d_m via the histogram threshold to calculate the boundary region of uncertainty. One of the methods used for obtaining threshold value in a continuous interval is the histogram threshold [49]. This method is used as a classification method for the number of two classes, whose objective is to reduce the ambiguity within the interval of the existing values.

For this work, first, the available unique values in the d_M matrix are collected; then, these values are divided into some intervals and the average of each interval is specified as (D_i) . In

the next step, the numbers of data pairs in each interval are counted ($g(D_i)$). Next, a weighted moving average of the window number of 5, $f(D_i)$, in equation (7) is calculated from these values. According to $f(D_i)$, we begin from the first

intervals and consider the first relative minimum $f(D_v)$ as the threshold value; in this way, the boundary to the uncertainty region is calculated according to equation (8):

$$f(D_i) = \frac{g(D_i)}{\sum_{e=1}^{z-1} g(D_e)} \times \frac{g(D_{i-2}) + g(D_{i-1}) + g(D_i) + g(D_{i+1}) + g(D_{i+2})}{5}, \quad \forall i = 2, 3, \dots, z-3. \quad (7)$$

$f(D_{v-1}) > f(D_v)$ and $f(D_v) < f(D_{v+1})$. Find the first valley points in the modified histogram:

$$m_d = D_v,$$

$$m_c = \max(D_i) - m_d,$$

$$\text{in d_M: } \begin{cases} \text{if } m_d \leq \text{two point distance} \leq m_c : \text{uncertainty region,} \\ \text{else : strong region.} \end{cases}$$

(8)

Then, two types of probability are calculated for belonging to neighborhoods. A similarity value between each data and each neighborhood is calculated from equation (9) based on S_SM matrix. Also, in equation (10), the degree of similarity is measured. If the value of the data pair in the d_m matrix is between the boundaries within the uncertainty region, we consider the value 0.1 for S :

$$\text{sim_toneigh}(x_i, N_h) = \frac{(1/|N_h|) \sum_{x_j \in N_h} S_SM_{ij}}{\sum_{z=1}^k (1/|N_z|) \sum_{x_j \in N_z} S_SM_{ij}}, \quad (9)$$

$$\text{degree_toneigh}(x_i, N_h) = \frac{(1/|N_h|) \sum_{x_j \in N_h} S(d_M_{ij})}{\sum_{z=1}^k (1/|N_z|) \sum_{x_j \in N_z} S(d_M_{ij})}, \quad (10)$$

$$S(d_M_{ij}) = \begin{cases} \text{if } m_d < d_M_{ij} < c_d : 0.1, \\ \text{else : } 0.9. \end{cases} \quad (11)$$

Then, to measure the level of uncertainty for each data for belonging to neighborhoods, we use equation (14). Generally, there are many criteria such as entropy for measuring the level of uncertainty, where selecting each of them does not affect the performance of our method.

We use a criterion based on the well-known silhouette index, which is often used in internal cluster validation. In

this method, for each data, the first highest probability (fm) and second highest probabilities (sm) of belonging to neighborhoods are selected and the level of uncertainty is determined. Finally, the method is applied to probabilities and then smoothed with coefficient β . We use $\beta=0.4$ which offers the best result. Equation (14) presents the method [24]:

$$\text{uncert_sim}(x_i) = \frac{2 * \text{sim_toneigh}(x_i, N_{fm})}{\text{sim_toneigh}(x_i, N_{fm}) + \text{sim_toneigh}(x_i, N_{sm})} - 1, \quad (12)$$

$$\text{uncert_degree}(x_i) = \frac{2 * \text{degree_toneigh}(x_i, N_{fm})}{\text{degree_toneigh}(x_i, N_{fm}) + \text{degree_toneigh}(x_i, N_{sm})} - 1, \quad (13)$$

$$\text{uncert}(x_i) = \beta * \text{uncert_sim}(x_i) + (1 - \beta) * \text{uncert_degree}(x_i). \quad (14)$$

Data with a smaller value in equation (13) (indicative of a greater level of uncertainty) is selected as the most informative data. Unlike other methods, we use local selection, instead of a global selection of informative data. In global methods, informative data is selected from all datasets. In the local method

in this paper, in order to balance the number of data in each neighborhood and, consequently, to balance clustering, the most informative data are first selected from the current pairwise constrained cluster, which has the maximum number of data. Consideration of the local method makes the results

stable and more accurate. Finally, the informative data are selected from the data that is not a member of neighborhoods.

4. Experimental Method

In this section, we empirically evaluate the performance and accuracy of the proposed method in comparison with the methods explained in related works. First, we explain the experimental setup and then the experimental results.

4.1. Dataset. There are three document clustering datasets that are normally used by many types of research: News-group20, Sector Dataset, and Webkb Dataset (<http://people.cs.umass.edu/~mccallum/data.html>). In order to reveal the robustness of our algorithm against different situations, five datasets with different classes and sizes are selected randomly from the three mentioned major datasets.

The fifth dataset which is represented with details in Table 1 is randomly selected from the third main dataset.

4.2. Evaluation Criterion. There are many criteria for evaluating document clustering. In this paper, two methods are used for the evaluation. Rand index (RI) is used for calculating the agreement between the labels obtained from the results of clustering compared to the class of real labels.

RI measures the agreement between two partitions, p_1 and p_2 , of the same dataset D . Each partition is viewed as a collection (in this case, ML and CL) of $n * (n - 1)/2$ pairwise decisions, where n is the size of D . For each pair of points d_i and d_j in D , P_i assigns them to either the same cluster or to different clusters. Let a be the number of decisions, where d_i and d_j are in the same cluster. Let b be the number of decisions where the two instances are placed in different clusters in both partitions. The total agreement can then be calculated using

$$RI(p_1, p_2) = \frac{(a + b)}{n(n - 1)/2}. \quad (15)$$

The second method is called normalized mutual information (NMI). This method is used for the evaluation of the assigned clustering label compared to the real class label of data. NMI considers both real and assigned classes' labels from clustering as two random variables. Then, it measures the mutual information obtained from these two random variables and normalizes them within the interval between zero and one. If C is the random variable of the assigned class from clustering and K is the random variable of real class for data sample, then NMI is obtained using

$$NMI = \frac{2I(C; k)}{H(c) + H(k)}. \quad (16)$$

In this formula, $I(C; K) = H(C) - H(C|K)$ is the mutual information between two random variables C and K , where $H(C)$ is the entropy of variable C and $H(C|K)$ refers to the conditional entropy of variable C given K . In order to obtain the best result, for each dataset, the proposed algorithm is run 10 times; then, we represent the average result.

4.3. Experimental Methodology. In order to evaluate the proposed method, three perspectives have been considered. These perspectives can evaluate the proposed method in terms of different aspects.

First perspective:

In this perspective, five datasets and two mentioned criteria are considered to compare the proposed algorithm with famous and similar algorithms. Some of the utilized algorithms are not for document clustering; therefore, at first, the documents are converted into document-term and similarity-document-term matrices. Then, the algorithms are run on these matrices. The result of these runs can be observed in Table 2 and Figure 2. These similar algorithms include:

- (1) Random selection: in this algorithm, active learning, neighborhoods, and similar concepts are not used, and pairs are randomly selected and presented to Oracle for responding. This algorithm is usually used as the basic approach.
- (2) PCKmeans [36]
- (3) NPU ([26])
- (4) URASC ([38])

Second perspective:

One of the differences between the proposed algorithm and state-of-the-art algorithms is the exploration of neighborhoods in the first phase. In the formation of neighborhoods, selection of data with Cannot-link response from oracles is considered. Therefore, the farthest-first selection strategy and random selection are usually used in the first phase. If the neighborhoods are completed earlier, a greater number of questions remain for the second phase; therefore, in the second phase, more informative data are selected, where more balanced neighborhood is consolidated; in this way, the accuracy and efficiency of the algorithm are enhanced. The number of questions from Oracle for the exploration of neighborhoods as well as also the accuracy and efficiency of the proposed algorithm in the first phase compared to the PCKmeans method can be observed in Table 3 and Figure 3.

Third perspective:

For showing the power of the semantic representation of the document, word2vector is used. So, for the evaluation, the proposed algorithm is implemented with and without word2vector representation. In case of without word2vector representation, the similarity matrix is obtained from the inverse to the distance matrix. Figure 4 demonstrates the result from this method for two datasets.

5. Experimental Results and Discussion

Figure 2 and Table 2 presents the run of the algorithms in the first perspective. In Figure 2, the y -axis represents the resulting clustering performances for the first perspective (measured by RI) while the x -axis indicates the total number of queries from Oracle. As mentioned previously, each curve

TABLE 1: Details of dataset for evaluation.

Dataset	Sampling details	Size of sample	Number of the term (after reduce)
The 20 newsgroups collection	News-diff-3 (ND3): comp.graphics, rec.autos, sci.crypt	450 : 3*150	8611 (390)
	News-sim-5 (NS5): comp.graphics, comp.os.ms-windows.misc, comp.sys.ibm.pc.hardware, comp.sys.mac.hardware, comp.windows.x	500 : 5*100	9388(450)
	News-diff-6 (ND6): alt.atheism, comp.graphics, misc.forsale, rec.autos, sci.crypt, talk.politics.guns	750 : 3*100 + 3*150	1228 (592)
The webkb dataset	Webkb-diff-univ-4 (WD4): Cornell, Texas, Washington, Wisconsin	600 : 4*150	13010 (672)
The sector dataset	Sect-diff-ind-9 (SD9): chemical.manufacturing.industry, aerospace.and.defense.industry, audio.and.video.equipment.industry, food.processing.industry, oil.and.gas.integrated.industry, consumer.financial.services.industry, biotechnology.and.drugs.industry, communications.equipment.industry, air.courier.industry	540 : 9*60	10913 (615)

TABLE 2: NMI details for the proposed algorithm.

Query method	Dataset	Q = 0	Q = 15	Q = 30	Q = 45	Q = 60
Random	ND3	26.86	23.64	34.78	39.78	42.74
	NS5	10.08	10.30	11.29	13.67	15.87
	ND6	18.51	17.56	24.98	29.35	29.97
	WD4	42.00	41.78	44.16	49.88	53.67
	SD9	18.88	19.56	21.78	23.05	23.96
PCkmeans	ND3	26.86	24.65	59.95	71.05	73.16
	NS5	10.08	10.56	15.25	18.76	24.78
	ND6	18.51	25.55	30.98	45.54	50.81
	WD4	42.00	39.14	64.86	68.46	69.10
	SD9	18.88	21.45	23.08	25.87	29.77
NPU	ND3	26.86	50.18	69.76	71.23	74.31
	NS5	10.08	10.54	16.65	18.72	21.68
	ND6	18.51	50.45	30.12	49.78	55.56
	WD4	42.00	41.54	66.35	70.79	74.68
	SD9	18.88	23.89	26.42	27.31	35.70
SVBPC	ND3	26.89	47.10	85.70	87.34	89.46
	NS5	11.28	13.78	17.15	19.06	24.71
	ND6	18.56	63.44	76.23	82.27	92.86
	WD4	42.75	56.08	81.15	87.45	89.84
	SD9	18.08	25.74	30.45	37.62	48.98
URASC	ND3	27.21	44.14	77.02	79.51	80.31
	NS5	13.26	15.78	15.41	18.10	23.89
	ND6	18.42	62.12	71.03	82.31	85.56
	WD4	40.05	41.03	55.32	62.40	66.19
	SD9	19.20	22.05	27.45	30.82	39.02

shows the average RI of proposed and state-of-the-art methods across 10 independent runs.

At the beginning of the curves, it is observed that curves are approaching each other as their number of informative constrained pairwise is similar. In the middle of each curve, with approximately 20 queries, the curves separated, but after 20 queries, all of the methods except the random method had a significantly high RI. As an explanation, we find that approximately after 20 queries, the high informative pairwise constraint is selected converging sections A and B, which is quick in algorithm 1. For example, our method has quick and robust convergence than other methods which has been followed by the URASC and NPU

methods, respectively. The number of queries in the proposed method and the first phase is low. Therefore, a large number of the query was saved for the second phase. This is another reason stated in the second perspective.

Table 2 reports an analysis based on NMI criteria. In this table, NMI such as RI has the same style. When the number of queries is 15, NMI has a significant jump after which the NMI value grows progressively. In our proposed method, the growth of NMI is usually greater than that of other methods. The reason for this occurrence is the same with RI and suggests that this same style is repeated with some differences. The same style in RI and NMI results suggests the reliability and validity of our method. However, URASC

(a)

(b)

(c)

(d)

FIGURE 2: Continued.

FIGURE 2: RI details on the proposed algorithm.

TABLE 3: Problems and solutions for handling reduction of accuracy by increasing the number of pairwise constraints.

Problems	Solutions
Effect of the disarrangement of pairwise constraints in a presentation to Oracle for reducing accuracy	For this problem, we use the distance of data points from SVM hyperplane (in this case, we use interval [0 1], instead of 0 and 1 to estimate the uncertainty of data point carefully)
Selection of pair constraint with inappropriate uncertainty	For this problem, we define uncertainty region concepts (in each iteration, they are created automatically)
Unbalanced consolidation of neighborhoods in terms of number	For this problem, we use local selection instead of global selection (we select the maximum uncertainty data point from a cluster with the maximum number)

FIGURE 3: RI & NMI details on the first phase.

FIGURE 4: RI details on the proposed algorithm with and without word2vector.

is similar to our proposed algorithm but URASC has a complicated statistical computation. Note that URASC is hard to adopt for document clustering and the NMI value is low sometimes.

The most important point in the investigation of these results is the balance and stability of the proposed algorithm, such that, in the proposed algorithm, the accuracy of the result usually improves by increasing the number of pairwise constraints.

However, sometimes there are reductions of accuracy and efficiency in the other algorithms by increasing the number of pairwise constraints. The reasons for these problems and some solutions which were considered in the proposed algorithm are stated in Table 3.

These problems occur when the selected pairwise constraint (pwc1) (with fewer queries (q_1)) is not a subset of the selected pairwise constraints (pwc2) (with more previous queries (q_2)). Generally, if equation (17) is satisfied, the stability of the results will be satisfactory. For satisfaction of this equation, Table 3 must be fulfilled. This equation is empirically obtained by studying Figure 2 and Table 2, as well as through testing the pairwise constraint in each iteration with different queries:

$$\begin{aligned}
 &\text{number} - \text{of} (q_1) < \text{number} - \text{of} (q_2) < \dots < \text{number} - \text{of} (q_n), \\
 &\quad \Rightarrow \\
 &\text{set} - \text{of} (pw1) \subseteq \text{set} - \text{of} (pw2) \subseteq \dots \subseteq \text{set} - \text{of} (pwn).
 \end{aligned}
 \tag{17}$$

In order to explain the second perspective, we analyze Table 4 and Figure 3. In Figure 3, the y-axis shows the resulting clustering performances for a second perspective (left side as measured by RI and right side as measured by NMI) and the x-axis indicates the five mentioned datasets. Each triangle point reveals the results of RI and NMI of our method while the rectangle point indicates the results of RI and NMI of our method with “farthest-first strategy” in the

first phase. We use 30 queries which are the average of a query from the first perspective. As mentioned earlier, we use the average result across 10 independent runs.

As can be seen in Table 4, our proposed algorithm usually asks fewer questions from Oracle in the phase of neighborhoods’ exploration for each of the five datasets. This reflects the relative superiority of the proposed algorithm in the first phase. In order to investigate the accuracy and efficiency in terms of two stated criteria, we implement the proposed algorithm in the first phase with two strategies of the “farthest-first strategy” which has been used in the published works and our proposed strategy. For example, NS5, ND6, and SD9 have a low query number with our proposed strategy in the first phase.

Figure 3 indicates that each dataset with our proposed algorithm has high RI and NMI values. Our method enjoys greater reliability and validity compared to other methods over the dataset with a variety of sizes and class numbers. In particular, our proposed method has offered better results over the dataset with a great size and great class number.

The third perspective is an important aspect of the proposed method. Application of word2vector can preserve the semantics and structure of the document. In contrary to other datasets, document clustering results depend on the semantic representation due to its unstructured content. Therefore, it is necessary to use a semantic representation which has been rarely used in the other state-of-the-art methods. In contrary to traditional semantic representation, use of deep learning offers a better result.

In Figure 4, the y-axis reveals the clustering performances obtained for the third perspective (for two datasets), and the x-axis indicates the number of the queries. This figure shows that use of word2vector can obviously improve the results. Following half of the queries, due to the gradual

TABLE 4: Number of queries from oracles in the first phase.

1st phase	Dataset				
	ND3	NS5	ND6	WD4	SD9
With farthest-first strategy	$Q = 6$	$Q = 12$	$Q = 11$	$Q = 9$	$Q = 23$
With proposed strategy	$Q = 4$	$Q = 7$	$Q = 12$	$Q = 6$	$Q = 18$

improvement of similarity, the curve of SVBPC with word2vector is separated significantly.

In any case, overall, our proposed method (first and second phases together) has offered the best results plus a higher efficiency and accuracy. Indeed, the results indicate that the proposed algorithm enhances efficiency and accuracy and the balance of results along with elevation of the pairwise number.

For tuning coefficients such as α in Algorithm 1, β in equation (15), and value of $\mathbb{S}$ operator in equation (13), the proposed algorithm runs with different values several times with main datasets, after which the best coefficients are selected.

α in Algorithm 1: this coefficient is for smoothing the center of current pairwise clusters and the center of neighborhoods. Indeed, we want to smooth Oracle response (in the form of neighborhoods) and the current pairwise clustering. At the beginning of the run, we find that a large α is better, but after the middle of the run, the small value is best as neighborhoods are not complete and they gradually would become complete and stable. Finally, we use $\alpha = 0.3$ since our method converges earlier, and this value offers a better accuracy.

β in equation (15): two types of probability, similarity, and degree of similarity are calculated for linking data points to neighborhoods. We want to determine the contribution of the degree of similarity to obtaining the level of uncertainty. If $\beta = 1$, it means that we do not use the degree of similarity, and our method would be unstable; otherwise, $\beta = 0$; then, we do not use the similarity probability. At the beginning of the run, great β is better as the neighborhoods are not completed. For efficiency at the beginning of the run, we use $\beta = 0.8$; then, (half of the number of queries) we use $\beta = 0.4$.

$\mathbb{S}$ operator in equation (13): the value of this operator is required for obtaining the degree of similarity in equation (13). Based on the values of the d_m matrix, we want to determine the weak pairwise relationship between data, where the values of the matrix are divided into three sections with histogram threshold. The values in the middle section involve a weak relationship between data. We want to assign the value of the operator to pairwise in the middle section. At the beginning of the run, a small value of the operator did not yield the best result as neighborhoods were not completed. Generally, we use a small value (approximately 0.1) for the value of the operator in the middle section. For two other sections, with a strong relationship, we use a large value for the operator (approximately 0.9). We can obtain stability of accuracy and performance with these values for the operator. However, it is better to use variable values for the operator in all sections though it is time consuming and costly.

6. Conclusion and Future Works

In this paper, first document-term and similarity-document-term matrices were made from documents; then, in an

iterative process, data with high uncertainty were selected for belonging to neighborhoods. In order to reach informative data, concepts of the SVM model, word2vec, neighborhoods, and uncertainty were used in each iteration. The proposed method outperformed the state-of-the-art methods with fewer queries from Oracle. In the phase of exploration, it achieved a better result with fewer questions. In the second phase, use of our proposed strategy alongside the uncertainty region balanced the number of data in each neighborhood. Generally, the obtained results were more balanced, which means that, by increasing the number of pairwise constraints, the accuracy grows simultaneously. The reason for the balance has been the uncertainty region and determining the degree of similarity for linking each data to the neighborhood. Using the SVM model, initialization of the current clustering based on pairwise constraints and updating of the centroid in each iteration and semantic representation yields considerable improvement of accuracy.

In future work, heuristic methods can be used to find the parameters of the proposed method. Deep learning tools can improve the results with semantic representation. For extending this research, one can use deep learning tools in any step of document constrained clustering especially in the similarity matrix and dimension reduction. In addition, hybrid methods can be used in each iteration, instead of using a support vector machine.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors received no financial support for the research and/or authorship of this article.

References

- [1] R. Sheikhpour, M. A. Sarraam, S. Gharaghani, and M. A. Z. Chahooki, "A survey on semi-supervised feature selection methods," *Pattern Recognition*, vol. 64, pp. 141–158, 2017.
- [2] A. Adolfsson, M. Ackerman, and N. C. Brownstein, "To cluster, or not to cluster: an analysis of clusterability methods," *Pattern Recognition*, vol. 88, pp. 13–26, 2019.
- [3] C.-L. Liu, W.-H. Hsiao, C.-H. Lee, and C.-H. Chen, "Clustering tagged documents with labeled and unlabeled

- documents,” *Information Processing & Management*, vol. 49, no. 3, pp. 596–606, 2013.
- [4] D. Vallejo-Huanga, P. Morillo, and C. Ferri, “Semi-supervised clustering algorithms for grouping scientific articles,” *Procedia Computer Science*, vol. 108, pp. 325–334, 2017.
 - [5] H. Huang and Y. Ma, “A hybrid clustering approach for bag-of-words image categorization,” *Mathematical Problems in Engineering*, vol. 2019, Article ID 4275720, 11 pages, 2019.
 - [6] R. Huang, W. Lam, and Z. Zhang, “Active learning of constraints for semi-supervised text clustering,” in *Proceedings of the 7th SIAM International Conference on Data Mining*, pp. 113–124, Minneapolis, MN, USA, April 2007.
 - [7] R. Huang and W. Lam, “Semi-supervised document clustering via active learning with pairwise constraints,” in *Proceedings of the IEEE International Conference on Data Mining, ICDM*, pp. 517–522, Omaha, NE, USA, October 2007.
 - [8] F. Gao, T. Huang, J. Sun, A. Hussain, E. Yang, and H. Zhou, “A novel semi-supervised learning method based on fast search and density peaks,” *Complexity*, vol. 2019, Article ID 6876173, 23 pages, 2019.
 - [9] A. K. Jain, “Data clustering: 50 years beyond K-means,” *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.
 - [10] X. Mai, J. Cheng, and S. Wang, “Research on semi supervised K-means clustering algorithm in data mining,” *Cluster Computing*, vol. 22, no. 2, pp. 3513–3520, 2019.
 - [11] B. Settles, *Active Learning Literature Survey*, University of Wisconsin-Madison Department of Computer Sciences, Madison, WI, USA, 2009.
 - [12] V.-V. Vu, H.-Q. Do, V.-T. Dang, and N.-T. Do, “An efficient density-based clustering with side information and active learning: a case study for facial expression recognition task,” *Intelligent Data Analysis*, vol. 23, no. 1, pp. 227–240, 2019.
 - [13] H. Zhang, S. Basu, and I. Davidson, “A framework for deep constrained clustering—algorithms and advances,” 2019, <https://arxiv.org/abs/1901.10061>.
 - [14] L. Cai, T. Yu, T. He, L. Chen, and M. Lin, “Active learning method for constraint-based clustering algorithms,” in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9659, pp. 319–329, Springer International Publishing, Cham, Switzerland, 2016.
 - [15] T. Van Craenendonck and H. Blockeel, “Constraint-based clustering selection,” *Machine Learning*, vol. 106, no. 9-10, pp. 1497–1521, 2017.
 - [16] D. R. Easterling, L. T. Watson, and N. Ramakrishnan, “Probability-one homotopy methods for constrained clustering,” *Journal of Computational and Applied Mathematics*, vol. 343, pp. 602–618, 2018.
 - [17] 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—ICML’04*, pp. 81–88, Banff, Canada, July 2004.
 - [18] A. Biswas and D. Jacobs, “Active image clustering with pairwise constraints from humans,” *International Journal of Computer Vision*, vol. 108, no. 1-2, pp. 133–147, 2014.
 - [19] D. Dinler and M. K. Tural, “A survey of constrained clustering,” in *Unsupervised Learning Algorithms*, pp. 207–235, Springer International Publishing, Cham, Switzerland, 2016.
 - [20] K. Pichara and A. Soto, “Active learning and subspace clustering for anomaly detection,” *Intelligent Data Analysis*, vol. 15, no. 2, pp. 151–171, 2011.
 - [21] S. C. H. Hoi, J. Zhu, R. Jin, and M. R. Lyu, “Semi-supervised SVM batch mode active learning for image retrieval,” in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, Anchorage, AK, USA, June 2008.
 - [22] Y. Leng, X. Xu, and G. Qi, “Combining active learning and semi-supervised learning to construct SVM classifier,” *Knowledge-Based Systems*, vol. 44, pp. 121–131, 2013.
 - [23] Q. Xu and K. L. Wagstaff, “Active constrained clustering by examining spectral eigenvectors,” in *Proceedings of the International Conference on Discovery Science*, Springer, Singapore, October 2005.
 - [24] D. Greene and P. Cunningham, “Constraint selection by committee: an ensemble approach to identifying informative constraints for semi-supervised clustering,” in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 4701, pp. 140–151, Springer Berlin Heidelberg, Heidelberg, Germany, 2007.
 - [25] T. Semertzidis, D. Rafailidis, M. G. Strintzis, and P. Daras, “Large-scale spectral clustering based on pairwise constraints,” *Information Processing & Management*, vol. 51, no. 5, pp. 616–624, 2015.
 - [26] S. Xiong, J. Azimi, and X. Z. Fern, “Active learning of constraints for semi-supervised clustering,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 1, pp. 43–54, 2014.
 - [27] S. S. Sonawane and P. A. Kulkarni, “Graph based representation and analysis of text document: a survey of techniques,” *International Journal of Computer Applications*, vol. 96, no. 19, pp. 1–8, 2014.
 - [28] M. M. Fazal and M. Rafi, “A semi-supervised approach to document clustering with sequence constraints,” *Journal of Independent Studies and Research—Computing*, vol. 13, no. 1, 2015.
 - [29] W. Zhou, H. Wang, H. Sun, and T. Sun, “A method of short text representation based on the feature probability embedded vector,” *Sensors*, vol. 19, no. 17, p. 3728, 2019.
 - [30] Z. Gao, Y. Li, Y. Sun et al., “Motion tracking of the carotid artery wall from ultrasound image sequences: a nonlinear state-space approach,” *IEEE Transactions on Medical Imaging*, vol. 37, no. 1, pp. 273–283, 2018.
 - [31] L.-X. Luo, “Network text sentiment analysis method combining LDA text representation and GRU-CNN,” *Personal and Ubiquitous Computing*, vol. 23, no. 3-4, pp. 405–412, 2019.
 - [32] T. Mikolov, K. Chen, G. Corrado, and J. Dean, “Efficient estimation of word representations in vector space,” in *Proceedings of the 1st International Conference on Learning Representations ICLR 2013*, Scottsdale, AZ, USA, May 2013.
 - [33] D. Eler, D. Grosa, I. Pola, R. Garcia, R. Correia, and J. Teixeira, “Analysis of document pre-processing effects in text and opinion mining,” *Information*, vol. 9, no. 4, p. 100, 2018.
 - [34] B. Altinel and M. C. Ganiz, “Semantic text classification: a survey of past and recent advances,” *Information Processing & Management*, vol. 54, no. 6, pp. 1129–1153, 2018.
 - [35] H. Guo and W. Wang, “An active learning-based SVM multi-class classification model,” *Pattern Recognition*, vol. 48, no. 5, pp. 1577–1597, 2015.
 - [36] 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*, pp. 333–344, Lake Buena Vista, FL, USA, April 2004.
 - [37] Y. Guo and D. Schuurmans, “Discriminative batch mode active learning,” in *Proceedings of the 20th International Conference on Neural Information Processing Systems*, Curran Associates Inc., Vancouver, Canada, pp. 593–600, 2007.

- [38] C. Xiong, D. M. Johnson, and J. J. Corso, "Active clustering with model-based uncertainty reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 1, pp. 5–17, 2017.
- [39] R. M. D. Oliveira, A. A. Chaves, and L. A. N. Lorena, "A comparison of two hybrid methods for constrained clustering problems," *Applied Soft Computing*, vol. 54, pp. 256–266, 2017.
- [40] F. Yang, T. Li, Q. Zhou, and H. Xiao, "Cluster ensemble selection with constraints," *Neurocomputing*, vol. 235, pp. 59–70, 2017.
- [41] S. Wei, Z. Li, and C. Zhang, "Combined constraint-based with metric-based in semi-supervised clustering ensemble," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 7, pp. 1085–1100, 2018.
- [42] Z. Yu, P. Luo, J. Liu et al., "Semi-supervised ensemble clustering based on selected constraint projection," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 12, pp. 2394–2407, 2018.
- [43] A. A. Abin and H. Beigy, "Active selection of clustering constraints: a sequential approach," *Pattern Recognition*, vol. 47, no. 3, pp. 1443–1458, 2014.
- [44] X. Xu and P. He, "Improving clustering with constrained communities," *Neurocomputing*, vol. 188, pp. 239–252, 2016.
- [45] M. Soleymani Baghshah, F. Afsari, S. Bagheri Shouraki, and E. Eslami, "Scalable semi-supervised clustering by spectral kernel learning," *Pattern Recognition Letters*, vol. 45, no. 1, pp. 161–171, 2014.
- [46] Z. Yu, "Incremental semi-supervised clustering ensemble for high dimensional data clustering," in *Proceedings of the 2016 IEEE 32nd International Conference on Data Engineering, ICDE 2016*, pp. 1484–1485, Helsinki, Finland, May 2016.
- [47] R. Lakshmi and S. Baskar, "DIC-DOC-K-means: dissimilarity-based initial centroid selection for document clustering using K-means for improving the effectiveness of text document clustering," *Journal of Information Science*, vol. 45, no. 6, pp. 818–832, 2019.
- [48] W. Zhao, Q. He, H. Ma, and Z. Shi, "Effective semi-supervised document clustering via active learning with instance-level constraints," *Knowledge and Information Systems*, vol. 30, no. 3, pp. 569–587, 2012.
- [49] T. Basu and C. A. Murthy, "A similarity assessment technique for effective grouping of documents," *Information Sciences*, vol. 311, pp. 149–162, 2015.