A Novel Text Ensemble Clustering Based on Weighted Entropy Filtering Model

Qiaoyun Shen1*, Yican Qiu2

1 Ningbo Polytechnic College, Ningbo, Zhejiang, 315800, China
2 Xi’an Jiaotong University, Xi’an, Shanxi, 710049, China
*Corresponding author’s e-mail: qy0523@126.com

Abstract. Text clustering is one of the important technical bases of natural language processing, and ensemble clustering improves the robustness of text clustering. According to the existing research of scholars and experts, the quality and diversity of basic clustering have a great influence on consensus clustering, and it has a particularly significant effect on text clustering. However, there are a few pieces of research aiming at reducing the number of low-quality clustering in ensembles. This paper proposes a novel clustering filtering model based on entropy criteria. The entropy criterion is used to evaluate the uncertainty of each cluster w.r.t. the ensemble. Two indexes are proposed on the basis of the uncertainty of cluster, namely, Clustering Trend Index (CTI) which indicates the contribution of each cluster w.r.t. basic clustering, and Cluster Consistency Index(CCI) which indicates the degree of cluster dispersion in the basic clustering. The proposed clustering filtering model is built on the basis of new weight using two proposed indexes. Thereby, by dropping the low-quality clustering, the percentage of high-quality clustering will increase. A large number of experiments on various real text data sets using optimal thresholds show that the proposed method has greatly improved accuracy and robustness, and is superior to existing ensemble clustering algorithms.

1. Introduction
Text clustering is an important step in natural language processing (NLP)[1], and text clustering can improve the accuracy of text processing. The quality of text clustering has a significant impact on the validity and accuracy of text analysis. Text clustering is used to categorize a set of large text documents into a subset of consonant and colorant clusters[2]. The essence of text clustering is the process of automatically categorizing document collections. The typical steps are to convert different documents from the original natural language text information into a data feature matrix with document features. And then calculate the similarity between the matrix rows and columns to achieve document clustering. A good text clustering makes the document content similarity in the same cluster as large as possible, and the document content similarity between different clusters as small as possible. So that the text information can be fully distinguished.

Text clustering is one of the hot topics in the field of NLP. The relevant research results are very rich. The main research contents of text clustering include text feature extraction and efficient text clustering. The main results of document feature extraction direction are [3], [4], [5], [6]. The main results of efficient document clustering direction are [7], [8], [9]. The above researches mainly propose or optimize an effective single clustering algorithm. However, according to the research of clustering problems, the performance of a single clustering algorithm is unstable. It is expressed that
different clustering algorithms are used for the same data set, but the clustering results are different. Even when the same clustering algorithm runs multiple times, the results are different.

In recent years, with more attention on ensemble clustering, a series of novel and efficient ensemble clustering algorithms have emerged, such as [10][11][12] [13][14] [15]. The main idea of ensemble clustering is to improve clustering performance by using multiple clustering to obtain a consensus result. Strehl et al. [10] proposed ensemble clustering firstly. In [10], ensemble clustering is divided into two basic directions: the generation of basic clustering and the integration of basic clustering, and a series of methods i.e. graph theory-based, hyper-graph partitioning method, association matrix based hierarchical clustering method are generated. Typically, the hierarchical clustering method based on the correlation matrix is common and efficient, and a typical method named EAC proposed in [11] constructs a co-association(CA) matrix [11] between data set elements through cluster tags of basic clustering, thereby avoiding the corresponding problem of cluster tags. At the same time, the traditional hierarchical clustering method is used to solve the CA matrix, and efficient clustering results are obtained. To solve the weight of basic clustering members, some weighted ensemble clustering methods have recently been proposed. Li and Ding [12] transformed the clustering problem into a non-negative matrix factorization problem and proposed a weighted consensus clustering method. In this method, each base cluster is assigned a weight to improve the consensus results. Huang et al. [13] suggested clustering based on the standardized group consistency index (NCAI) and weight basis. Huang Dong proposed LWCA [14] based on the EAC method. This method is based on a locally weighted strategy and obtained a better clustering effect than the former. To distinguish the effects of different basic clustering on the clustering results, Liang et al. [15] proposed a method for measuring the quality of basic clustering members based on the three-way decisions idea, and built the filtering method based on the three-way decision model and improved the performance of the ensemble algorithm.

The above research shows that ensemble clustering improves the performance of clustering, and it is an innovative attempt to introduce the ensemble clustering method into the field of text clustering. In the process of text clustering, different features must express the clustering results differently, so reasonable feature extraction and clustering selection will directly affect the quality of the clustering results. To improve the positive impact of basic clustering with more useful features, can greatly obtain a better consensus result than classical algorithms. Because the quality of basic clustering has an important influence on the effect of ensemble clustering, it is also a research idea to improve the effect of ensemble clustering to screen and optimize basic text clustering. At the same time, some scholars have researched this issue, but they have not applied it to the field of text clustering. In this paper, we propose a novel basic clustering filtering model to improve the percentage of high-quality basic clustering, to obtain a better text ensemble clustering method.

The rest of the paper is organized as follows. The related work is reviewed in Section II. The proposed model based on entropy theory and novel ensemble clustering based on the proposed model is described in Section III. The experimental results are reported in Section IV. And in the end, we conclude the paper in Section V.

2. Related works

2.1. Text clustering
The general process of text clustering has three steps: text pre-processing, feature space dimension reduction, and clustering using features[4].

The first step is text pre-processing, which is to segment the text and extract keywords. This step consists of four parts, which are tokenization, stop words removal, stemming, and term weighting. A lot of pre-processing methods or models are generated in this step. Among them, the vector space model (VSM)[16] is used by the majority of text document paradigms in text clustering to weigh each document. In this model, each term present in the text documents is featured for document representation [17]. The text documents are represented by a multi-dimensional space, in which the
position value of each dimension corresponds to a weight value. The text features generated from
different text terms, even small documents in a collection, would be represented in hundreds and
thousands of text features.

In the second step, the feature space dimension reduction and the feature vector space are processed
using a suitable feature extraction method or feature extraction. Some feature selection or feature
extraction methods such as mutual information [18], sequential search algorithms [19], principal
component analysis (PCA) [20], and Neural networks [21] text clustering are used to reduce the
uninformative feature. And then obtaining a feature space with much more useful information and
suitable dimension.

In the third step, clustering with the processed features. That is, the vectorized data in feature space
which is reduced in step 2 is clustered by using a text clustering algorithm. After the vector space
model is created, the vector points in the vector space model are clustered by using the determined
clustering algorithm. Figure 1 is the process of text clustering. The process in Fig. 1 can also be used
as the basic clustering part in ensemble clustering.

2.2. Ensemble clustering
Ensemble clustering is mainly divided into the generation of basic clustering and the integration of
basic clustering(Figure 2).

The generation of basic clustering is a process of clustering data set \( D \) many times. Among them,
the set consisting of basic clustering produced by multiple clustering algorithms is called ensemble \( \Pi \)
where \( \pi_i \) represents the \( i \)-th clustering result in ensemble \( \Pi \), and \( M \) means there are \( M \) basic clustering
members in ensemble \( \Pi \). Each basic clustering member is a clustering result of the data set \( D \). Each
basic clustering member can be obtained by different clustering algorithms of simple clustering or by
setting different initial parameters of a single clustering algorithm. Wherein each basic clustering
member comprises a plurality of clusters and is recorded as \( \pi_i = \{c_{i1}, c_{i2}, \ldots, c_{in} \} \), where \( c_{ij} \) represents the
\( j \)-th cluster in basic clustering \( \pi_i \), and \( n_i \) means there are \( n_i \) cluster members in basic clustering \( \pi_i \).
Hence, according to the basic definition of ensemble clustering, a set contains \( M \) basic clustering
members can be recorded as \( \Pi = \{\pi_1, \pi_2, \ldots, \pi_M \} \).
The integration of basic clustering is to aggregate the set of basic clustering with appropriate ensemble methods. Usually, the ensemble methods include the hierarchical clustering method [22], voting method [23], information theory method [24], and graph partitioning method, etc. Among them, the hierarchical clustering method is widely used because it is concise and efficient, and makes better use of the correlation matrix (such as CA matrix) between basic clusters, which can obtain a better clustering effect. Suppose that the times of any two elements \( p_i \) and \( p_j \) in the text data set Document = \{ \( p_1, p_2, \ldots, p_n \) \} are clustered together in ensemble \( \Pi \) is \( X_c \), then the probability of these two elements appearing in the same cluster is computed as \( X_c / M \). Therefore, the relationship matrix called the CA matrix between the basic clusters can be obtained by calculation. For a given text data set Document = \{ \( p_1, p_2, \ldots, p_n \) \}, an \( n \times n \) dimension matrix based on the elements of the text data set, i.e. CA matrix, can be constructed and computed as:

\[
CA(\Pi) = \left\{ \frac{\sum_{i,j} \delta_{ij}}{M} \right\}_{n \times n}
\]

(1)

\[
\delta_{ij} = \begin{cases} 
1 & \text{if } \text{Cls}(p_i) = \text{Cls}(p_j) \\
0 & \text{otherwise}
\end{cases}
\]

(2)

Wherein \( \text{Cls}(t_i) \) represents which cluster the sample object belongs to in the M-th basic clustering. So far, the CA matrix of text ensemble clustering is constructed, and the consensus clustering result is obtained by using hierarchical clustering or other methods to solve the CA matrix.

3. Text Ensemble Clustering Based on Weighted Entropy Filtering Model

Although the ensemble strategy improves the robustness of text clustering, it results in the superiority and inferiority of basic clustering affecting the clustering results. At the same time, there are many non-information features in the text language itself, and the expression of clustering results by these features is necessarily different. To further reduce the impact of basic clustering itself on consensus clustering results and to enhance the positive impact of high-quality basic clustering on consensus clustering results, this section will propose a text-based clustering filtering model based on information entropy and a novel ensemble clustering method on basis of this model.

3.1. Information entropy cluster uncertainty measure

Due to the nature of unsupervised learning, the association between a set of random variables is difficult to give directly. The information entropy in information theory happens to be a measure of the uncertainty associated with random variables. Information entropy is a commonly used indicator of uncertainty relevance. In information theory [24], information entropy is a measure of the uncertainty associated with random variables. The degree of influence of different basic clusters on clustering results can be given by the degree of association between clusters. The formal definition of information entropy is given in Definition 1.

**Definition 1** For a discrete random variable \( x \), the information entropy \( H(x) \) is defined as:

\[
H(x) = \sum_p p(x) \log_2 p(x) = \sum_p p(x) \log_2 \frac{1}{p(x)}
\]

(3)

Where \( p(x) \) is the probability mass function of variable \( x \).

The uncertainty of a set of random variables is usually measured by the joint entropy. Therefore, joint entropy is generated by extending the basic information entropy theory. The general definition of joint information entropy is as defined in Definition 2.

**Definition 2** For a set of discrete random variables \( \psi = (X_1, \ldots, X_r) \), the joint entropy is defined as:

\[
H(\psi) = \sum_{i=1}^r H(X_i) = H(X_1) + \cdots + H(X_r)
\]

(4)
For any basic text clustering $\pi \in \Pi$, that is, $n$ clusters in $\pi' = \{c_1', c_2', \ldots, c_n'\}$ are independent of each other. Therefore, for any cluster $C_i$ in $\Pi$ (that is, any cluster in any basic clustering), the uncertainty of $C_i$ relative to a certain basic text clustering $\pi'$ can be calculated as the sum of the probability that the elements in $C_i$ appear in each cluster of $\pi'$. 

**Definition 3** For any cluster $C_i$ in the ensemble $\Pi$, since the clusters in the same basic clustering are all independent, the uncertainty compared with a certain cluster $\pi_i$ is computed by the joint entropy in definition 2, and the calculation formula is as follows:

$$H(C_i) = -\sum_{j=1}^{n'} H(C_j)$$

where $p(C_i, C_j') = \frac{|C_i \cap C_j'|}{|C_i|}$, $C_i$ is the probability that element in $C_i$ appears in each cluster of basic clustering, $n'$ is the number of clusters in $\pi'$, $c_j'$ is the $j$-th cluster in $\pi'$, and $|\cdot|$ outputs the number of objects in $\cdot$.

In the same way, because the basic clustering members are independent of each other (generated by multiple clustering algorithms), the premise that the random variables are independent of each other is satisfied. Therefore, the uncertainty of one basic clustering w.r.t the whole ensemble can be obtained by the concept of joint entropy, which is defined as:

**Definition 4** For any of the basic clustering $\pi_i$ in ensemble $\Pi$, due to the basic clustering are all independent, the uncertainty of $\pi_i$ compared to ensemble $\Pi$ is computed by the joint entropy in definition 2. The calculation formula is as follows:

$$H(\pi_i) = -\sum_{j=1}^{n'} H(C_j) = H(C_1') + \cdots + H(C_{n'}')$$

### 3.2. Entropy weighted basic clustering filtering model

Given the reliability weights of the various clusters already given in Definition 3, and considering different basic clustering, there may be different clusters numbers. And the quality of the basic clustering needs to consider the coordination between its clusters. Hence the quality of basic clustering is hard to be given through joint entropy w.r.t Definition 2. To solve this problem, we will present two new weighing indicators. That is, the CTI for measuring the overall trend of cluster reliability in clustering and the CCI for measuring the degree of uniformity between clusters in basic clustering. Therefore, the quality of a basic clustering can be defined by a two-tuple definition of its CTI and CCI indicators, defined as:

**Definition 5** For basic clustering $\pi'$ in ensemble $\Pi$, we agree to measure its pros and cons indicators as quality weights named $W(\pi')$, and its quality weight is defined as a two-tuple. Therefore, the quality weights are computed as:

$$W(\pi') = \langle CTI(\pi'), CCI(\pi') \rangle$$

Wherein, the CTI is defined as the contribution of clusters in $\pi'$. And the CTI is computed as:

$$CTI(\pi') = \frac{\sum_{j=1}^{n'} H(C_j')}{n'}$$

the CCI is defined as the degree of discretization of clusters in $\pi'$. And the CCI is computed as:

$$CCI(\pi') = \frac{\sqrt{\sum_{j=1}^{n'} (H(C_j') - CTI(\pi'))^2}}{n'}$$

where $n'$ represents the number of clusters in $\pi'$.

At this point, the quality weight of the basic clustering can be obtained by Definition 5. And then, we can build a basic clustering filtering model based on the quality weights. The model named Entropy weighted basic clustering filtering model is constructed as:
For a given CTI threshold $\alpha$ and CCI threshold $\beta$, determining whether the basic clustering is reliable on basis of its quality weight, and then according to the following rules to construct the three-domain sets of the proposed model (the three-domain sets are POS, NEG, and BND):

1) if $\text{CTI}(\pi_i) \geq \alpha$ and $\text{CCI}(\pi_i) \geq \beta$, then the basic clustering $\pi_i$ is reliable and put the $\pi_i$ into the set of POS. That is $\pi_i \in \text{POS}$.

2) if $\text{CTI}(\pi_i) \leq \alpha$ and $\text{CCI}(\pi_i) \leq \beta$, then the basic clustering $\pi_i$ is unreliable and put the $\pi_i$ into the set of NEG. That is $\pi_i \in \text{NEG}$.

3) if it does not satisfy the conditions 1) and 2), then the basic clustering $\pi_i$ will be put into the set of BND. That is $\pi_i \in \text{BND}$.

The objects in the sets of POS are all clustering which has high quality. Conversely, The objects in the sets of NEG are all clustering which has low quality. And the objects in BND are all clustering whose quality is unstable and this clustering may translate into the object in POS or NEG. So, different decision behaviors are used for objects in three different decision domains. To improve the reliability of clustering integration by the overall basic clustering, the algorithm model will be iteratively executed by introducing the strategy of discarding the elements in the set of NEG, preserving the elements in the set of POS, and promoting the transformation of the elements in BND to the other domain sets. With the proportion of the quality basic clustering reaching 100% or the iteration reaching $R$ times, the filtering model will stop.

For clarity, the overall algorithm of the filtering model is summarized in Algorithm 1.

Algorithm 1: Weighted Entropy Filtering Model Text Ensemble Clustering

**Input:** $D$ (text data set), $k$ (number of consensus clustering), $R$ (times of iteration), thresholds $\alpha$ and $\beta$

1. Processing the text data set $D$ by using classic text processing methods.
2. Get the ensemble $\Pi$ by using text clustering algorithm (such as k-means) $M$ times.
3. Compute the weight of the base clustering in $\Pi$ w.r.t. Definition 1-5.

Construct the model iteratively:

4. for $t = 1, 2, ..., R$
   5.   for $i = 1, 2, ..., |\Pi|$ |
      6.     if $\text{CTI}(\pi_i) \geq \alpha$ and $\text{CCI}(\pi_i) \geq \beta$ then |
          Add $\pi_i$ into the set of POS |
      7.     else if $\text{CTI}(\pi_i) \leq \alpha$ and $\text{CCI}(\pi_i) \leq \beta$ then |
          Add $\pi_i$ into the set of NEG |
      8.     else |
          Add $\pi_i$ into the set of BND |
          end if |
      end for |
  9.   if $|\text{NEG}(\Pi)| = 0$ then |
  10.     Change the sets of POS and BND into $\Pi'$ |
          end if |
      end for |
  11. Delete the set of NEG and change the sets of POS and BND into $\Pi'$ |
      Ending iteration |
  12. Construct the co-association matrix $E$ using $\Pi'$ |
  13. Reduce the dimensionality of $E$ using EAC |
  14. Obtain the clustering with $k$ clusters in the $k$ dimensional matrix |

**Output:** Consensus clustering $\pi'$
4. Experiments
In this section, we evaluate the proposed methods against the several efficient ensemble clustering methods on a variety of real-world data sets. The clustering algorithm is evaluated by NMI[15]. The text data set used in our experiments can be obtained at the web of LABIC(http://sites.labic.icmc.usp.br/text_collections/). All experiments are conducted in pyCharm professional 2017.1 64-bit on a workstation (Ubuntu 18.04 LTS, Intel I7-7820X 4.0GHz processors, 32 GB of RAM).

4.1. Data sets and evaluation methods
In our experiments, eight real-world data sets are used, namely, Technical reports (DS1), Web pages (DS2), TREC with different documents (DS3, DS4, DS5, DS6), MEDLINE (DS7), and 20 newsgroup (DS8). All of the data sets are from LABIC(Laboratory of Computational Intelligence), and the details of the eight data sets are given in Table 1.

| Dataset   | Source          | * of documents | * of terms | * of clusters |
|-----------|-----------------|----------------|------------|---------------|
| DS1       | Technical reports | 299            | 1725       | 4             |
| DS2       | Web pages       | 333            | 4339       | 4             |
| DS3       | TREC            | 204            | 5832       | 4             |
| DS4       | TREC            | 313            | 5804       | 6             |
| DS5       | TREC            | 414            | 6429       | 8             |
| DS6       | TREC            | 878            | 7454       | 9             |
| DS7       | MEDLINE         | 913            | 3100       | 10            |
| DS8       | 20NEWGROUP      | 18828          | 45433      | 20            |

Normalized Mutual information (NMI)[15] is a common clustering evaluation criterion. The NMI measure provides a noise indication of the shared information between two clustering members. Let $\pi'$ be the test clustering and $\pi$ the ground-truth clustering. The NMI score of $\pi'$ w.r.t. $\pi$ is defined as follows:

$$NMI(\pi', \pi) = \frac{\sum_{i=1}^{c'} \sum_{j=1}^{c} n_i \log \frac{n_i n_j}{n_i n_j}}{\sum_{i=1}^{c} \sum_{j=1}^{c} n_i \log \frac{n_i}{n} \sum_{j=1}^{c} n_j \log \frac{n_j}{n}}$$

where $n'$ is the number of clusters in $\pi'$, $n$ is the number of clusters in $\pi$, $n_i$ is the number of objects in the $i$-th cluster of $\pi'$, $n_j$ is the number of objects in the $j$-th cluster of $\pi$, and $n_{ij}$ is the number of common objects shared by $i$-th cluster in $\pi'$ and $j$-th cluster in $\pi$. 


4.2. Design and preparation of experiments
Before the experiment begins, we use a classical text clustering algorithm (such as k-means) to generate 200 candidate basic clustering members named candidate clustering pool (CCP). In this work, a pool contains 200 candidate clustering members is randomly generated for each data set. Based on CCP, we will carry out several groups of comparative experiments, namely, the proposed algorithm compares with basic text clustering, the average performance of each ensemble clustering who runs data sets 100 rounds, the stability of each ensemble clustering on the size of ensemble $\Pi$, and the time complexity of several clustering ensembles. In addition, the proposed method which attends the comparative experiments is named PM. The benchmark ensemble clustering methods involved in the comparison are LWEA[14], LWGP[14], EAC, HGPA[10].

4.3. Choices of thresholds $\alpha$, $\beta$
In this section, the thresholds are determined by using the simulated annealing algorithm on different data sets $\alpha$ And threshold $\beta$. The experiment will be repeated 50 times, and the average of the results of 50 times will be plotted. The final experimental results are shown in Figure 3.

![Figure 3. Calculation of thresholds $\alpha$ and $\beta$.](image)

The $x$-axis of $xy$ plane is naturally growing $\alpha$, the $y$-axis is naturally growing $\beta$. The $z$-axis(verticall axis) is the promotion rate of the ensemble clustering performance compared with the unfiltered results under different thresholds. Because the threshold can adjust the ratio of high-quality basic clustering generated by the model, the consensus clustering has better performance. As shown in Figure 4, the promotion rate of the algorithm changes with the change of the threshold. Especially when $\alpha$ gets a value between 0.9 and 1.2, $\beta$ gets a value between 0.09 and 0.11, the promotion effect is better. When the threshold reaches a critical value, negative optimization will occur. In order to ensure the effect of follow-up experiments, the threshold of follow-up experiments $\alpha$ takes 1.1, $\beta$ takes 0.1.

4.4. Comparison against basic text clustering
Compared with the basic clustering method, ensemble clustering achieves more robust clustering results. Therefore, in this section, we firstly compare several ensemble clustering methods (EAC, LWEA, LWGP) with the basic clustering methods. For each benchmark data set, we choose 50 clustering members from CCP randomly, and then we run 100 rounds by using each ensemble clustering method. Finally, we calculate the average NMI score of 50 rounds as the score of each ensemble clustering method. Figure 4 shows the average NMI scores of various algorithms.
Figure 4. Average NMI scores of several ensemble clustering methods and basic clustering over 50 rounds.

Compared with the basic clustering algorithm (choosing 50 clustering members from CCP and computing the average NMI scores), the ensemble clustering algorithm has a significant improvement in stability. Especially, for DS1, DS2, DS4, DS6, DS8 data sets, the ensemble clustering method has more advantages than basic clustering. Compared with several clustering ensemble methods, the PM has a remarkable clustering effect on DS3, DS5, and DS8.

4.5. Comparison against other ensemble clustering methods
In this section, we compare the PM method with four ensemble clustering methods, namely LWEA, LWGP, HGPA, EAC. To compare and analyze several ensemble clustering algorithms, we use gt-k (ground truth-k) to specify the cluster number of consensus clustering results. For gt-k, each benchmark method uses the truth number of clusters in the data set. The result of the consensus that runs with different ensemble sizes is illustrated in Table 2. To make the experimental results unaffected by special circumstances, we have conducted 100 rounds of experiments for each method with ensemble size = 40 (the experimental results are shown in Table 3). By the way, to control the stability effect caused by ensemble size, five groups of ensemble size experiments were carried out on each benchmark data set. The average NMI scores of 100 rounds are used to express the performance of each method.

As shown in Table 2, the PM method obtains the best NMI scores on eight benchmark data sets. Among them, the NMI scores on DS1, DS5, DS6, and DS8 are more remarkable. Compared with the other two ensemble clustering algorithms (EAC and HGPA), PM method and LWEA and LWGP have achieved better results on DS2, DS7, and DS8 data sets. From the perspective of stability, when the ensemble size is small on DS1, DS3, and DS4, the performance of LWEA and PM is similar. However, with the increase of ensemble size, the PM method still performs better than LWEA. Compared with HGPA and EAC, the data sets of DS2, DS4, DS5, DS7, and DS8 have a large gap. In addition, on almost all data sets, the PM method achieves the best or better NMI score.
As shown in Table 2, The NMI score of LWEA in DS4 and DS7 is better than the PM method. However, as shown in Table 3, with the change of ensemble size, the PM method performs better than LWEA because of the filtering model giving a high-quality ensemble. In particular, as shown in Table 3, under a specific integration scale, the proposed algorithm PM gives full play to the advantages of the filtering mechanism. By analyzing the number of filtered-based clustering members in this scale,
low-quality base clustering members certainly harm the performance of the ensemble clustering algorithm. In addition, there is no fixed value interval for the filtering of text-based clustering members, that is, the number of filtered members may not be the key to the improvement of algorithm performance. In terms of stability, several methods all have a better performance than the base clustering algorithm. Among them, the \textbf{PM} method has the best stability for all data sets.

Table 3. Average NMI scores over 100 rounds of each method (the ensemble size is 40).

| Method | LWGP   | HGPA   | EAC    | LWEA   | PM       | $|\text{Neg}(\Pi)|$ |
|--------|--------|--------|--------|--------|----------|------------------|
| DS1    | 0.5635±0.012 | 0.5104±0.024 | 0.5282±0.025 | 0.5888±0.016 | 0.6151±0.014 | 15               |
| DS2    | 0.7919±0.018 | 0.6897±0.032 | 0.6855±0.028 | 0.7918±0.011 | 0.8091±0.009 | 25               |
| DS3    | 0.4503±0.018 | 0.3965±0.039 | 0.4419±0.038 | 0.4839±0.021 | 0.5166±0.022 | 23               |
| DS4    | 0.5999±0.016 | 0.5478±0.027 | 0.5145±0.025 | 0.6218±0.021 | 0.6290±0.018 | 22               |
| DS5    | 0.5501±0.016 | 0.5498±0.027 | 0.5423±0.027 | 0.5681±0.019 | 0.5945±0.016 | 31               |
| DS6    | 0.5945±0.017 | 0.5488±0.029 | 0.5658±0.028 | 0.6189±0.018 | 0.6401±0.015 | 18               |
| DS7    | 0.5310±0.015 | 0.4659±0.018 | 0.4803±0.019 | 0.5443±0.017 | 0.5401±0.018 | 9                |
| DS8    | 0.4855±0.015 | 0.3932±0.024 | 0.4205±0.027 | 0.4907±0.022 | 0.5284±0.018 | 12               |

The above experiment quantitatively proves the performance of \textbf{PM}, and this part will verify the impact of the ensemble size $M$ on all ensemble clustering algorithms. The evaluated result will give the performance of \textbf{PM} and several ensemble clustering methods by using ensemble size $M$ (the cardinal $M$ of different basic clustering sets). For each ensemble size $M$, we run the proposed methods and the baseline methods 20 times on each benchmark data set (DS1, DS2, DS3, and DS4), with the ensemble of $M$ basic clustering members randomly selected from CCP at each time. The average NMI scores are shown in Figure 5.

As shown in Figure 5, almost on all benchmark data sets, the ensemble size has the least impact on \textbf{PM} and LWGP. On benchmark data set, the ensemble size has the most impact on EAC and HGPA. In addition, with the increase of ensemble size, \textbf{PM} has much better performance in all benchmark data sets because of the filtering characteristics.
5. Conclusion

In this paper, to further improve the accuracy of text clustering in scene text, we have proposed a novel text clustering ensemble approach based on a weighted entropy filtering model. We propose two indexes to evaluate the quality of basic clustering and a quality weight of basic clustering. A basic clustering filtering model has been constructed based on the proposed quality weight. Thus, the impact of bad clustering on clustering results is greatly reduced. Then a three-step clustering ensemble method based on the weighted entropy filtering model is proposed. We have carried out extensive experiments on various real data sets. The experimental results show that this method is superior to the existing methods in clustering quality and efficiency.

Although the proposed algorithm achieves a better algorithm performance, the computational complexity of the algorithm is significantly higher than the traditional algorithms. Despite when the ensemble size is large, the proposed method is more stable and accurate. Therefore, the follow-up work will be based on algorithm optimization. In particular, with the introduction of the integration idea, integration optimization and algorithm complexity optimization will be the focus of the next work schedule.

Acknowledgments

This work is supported by the General fund of Zhejiang Education Department (No. Y201941153) and General project fund of Ningbo Polytechnic College (No. NZ19C009).

References

[1] Chowdhury, G.G. Natural language processing. Annual review of information science and technology 2003, 37, 51–89.
[2] Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A. Multi-objectives-based text clustering technique using K-mean algorithm. International Conference on Computer Science & Information Technology, 2016.

[3] Abualigah, L.M.; Khader, A.T.; Al-Betar, M.A.; Alomari, O.A. Text feature selection with a robust weight scheme and dynamic dimension reduction to text document clustering. Expert Systems with Applications An International Journal 2017, 84, 24–36.

[4] Kushwaha, N.; Pant, M. Link based BPSO for feature selection in big data text clustering. Future generation computer systems 2018, 82, 190–199.

[5] Vieira, A.S.; Borrajo, L.; Iglesias, E.L. Improving the text classification using clustering and a novel HMM to reduce the dimensionality. Computer Methods & Programs in Biomedicine 2016, 136, 119–130.

[6] Bharti, K.K.; Singh, P.K. A three-stage unsupervised dimension reduction method for text clustering. Journal of Computational Science 2014, 5, 156–169.

[7] Mikhina, E.K.; Trifalenko, V.I. Text clustering as graph community detection. Procedia Computer Science 2018, 123, 271–277.

[8] Tao, P.; Lu, L. A novel incremental conceptual hierarchical text clustering method using CFu-tree. Applied Soft Computing Journal 2015, 27, 269–278.16 of 16

[9] Zhang, W.; Yoshida, T.; Tang, X.; Wang, Q. Text clustering using frequent itemsets. Knowledge-Based Systems 2010, 23, 379–388.

[10] Strehl, A.; Ghosh, J. Cluster ensembles - A knowledge reuse framework for combining multiple partitions; Vol. 3, 2003; pp. 583–617.

[11] Cristofor, D.; Simovici, D. Finding Median Partitions Using Information-Theoretical-Based Genetic Algorithms. Journal of Universal Computer Science 2008, 8, 153–172.

[12] Zhang, Y.; Zeng, E.; Li, T.; Narasimhan, G. Weighted Consensus Clustering for Identifying Functional Modules In Protein-Protein Interaction Networks. 2009, pp. 539–544.

[13] Huang, D.; Lai, J.H.; Wang, C.D. Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis. Neurocomputing 2015, 170, 240–250.

[14] Huang, D.; Wang, C.D.; Lai, J.H. Locally Weighted Ensemble Clustering. IEEE Transactions on Cybernetics 2016, 48, 1460–1473.

[15] Liang, W.; Zhang, Y.; Xu, J.; Lin, D. Optimization of Basic Clustering for Ensemble Clustering: An Information-Theoretic Perspective. IEEE Access 2019, 7, 179048–179062.

[16] Seifzadeh, S.; Farahat, A.K.; Kamel, M.S.; Karray, F. Short-Text Clustering using Statistical Semantics 2015.

[17] Abualigah, L.M.; Khader, A.T. Unsupervised text feature selection technique based on hybrid particle swarm optimization algorithm with genetic operators for the text clustering. Journal of Supercomputing 2017, pp. 1–23.

[18] Alush, A.; Friedman, A.; Goldberger, J. Pairwise clustering based on the mutual-information criterion. Neurocomputing 2016, 182, 284–293.

[19] Revanasiddappa, M.B.; Harish, B.S.; Kumar, S.V.A. Meta-cognitive Neural Network based Sequential Learning Framework for Text Categorization. Procedia Computer Science 2018, 132, 1503–1511.

[20] Lam, S.L.Y.; Lee, D.L. Feature reduction for neural network based text categorization. International Conference on Database Systems for Advanced Applications, 1999.

[21] Kushwaha, N.; Pant, M. Textual data dimensionality reduction - a deep learning approach. Multimedia Tools and Applications 2018.

[22] Akbari, E.; Dahlan, H.M.; Ibrahim, R.; Alizadeh, H. Hierarchical cluster ensemble selection. Engineering Applications of Artificial Intelligence 2015, 39, 146–156.

[23] Zhou, Z.H.; Tang, W. Clusterer ensemble. Knowledge-Based Systems 2006, 19, 77–83.
[24] Wang, F.; Yang, C.; Lin, Z.; Li, Y.; Yuan, Y. Hybrid sampling on mutual information entropy-based clustering ensembles for optimizations. Neurocomputing 2010, 73, 1457–1464.