Study on Ensemble based Clustering Algorithm for Gene Expression Data

Zhenfang Chu¹, Buyang Cao² and Fang Yu¹

¹School of Software Engineering, Tongji University, Shanghai, China
²College of Architecture and Urban Planning, Tongji University, Shanghai, China.
Email: 1641470@tongji.edu.cn

Abstract. Cluster analysis, one of the most powerful tools for analysing gene expression data, has been frequently used to obtain information and knowledge for supporting decision-making for the diagnosis and treatment of diseases. Recently, more scholars are dedicated to ensemble-based clustering algorithms for gene datasets. The consensus function is the key of ensemble clustering to integrate clustering results of the base clustering algorithms. In this paper, we propose a new consensus function that transforms multiple clustering results into a co-association matrix according to the weighted voting derived from the mutual information, where the values determining whether two samples belong to the same cluster. The computational experiments on multi-dataset of gene expression using the ensemble clustering algorithm based the new consensus function are conducted. Comparing to the base clustering algorithms including K-Means, DBSCAN and hierarchical clustering as benchmarks, our algorithm presents its superiority.

1. Introduction

Gene expression data analysis gets more attentions these days [1]. For gene expression data analysis, especially in the absence of prior information on data, the clustering analysis is an important tool and widely applied since it is particularly suitable for unsupervised learning.

Ensemble based algorithm was originally designed for supervised learning, whose idea is to combine multiple base learners in an effective way to solve a learning task. The integration of the base learners can often lead to the improvement in learning performance and better solutions comparing to base learners. Nowadays, many researchers have being applied ensemble learning algorithms to gene-data related clustering problems in practice and reveals the advantages of ensemble clustering over the traditional base clustering algorithms.

The main idea of an ensemble clustering algorithm is to use multiple clustering algorithms as base learners to create the clusters on the dataset and then use the consensus function to integrate all resultant clusters into the final solution. Inspired by the advantages of an ensemble algorithm, we propose a consensus function that transforms multiple clustering results into a co-association matrix by weighted voting according to the mutual information. The consensus function provides the foundation of our ensemble clustering algorithm. A value in the co-association matrix represents whether two underlying data objects belong to the same cluster. The final clustering results are generated according to the matrix.

To evaluate the performance of the proposed ensemble clustering algorithm, we perform multiple computational experiments using K-Means, DBSCAN and hierarchical clustering as the benchmarks. Different evaluation metrics are chosen for the labelled and unlabelled datasets to evaluate the results.
The metrics include the adjusted rand index, adjusted mutual information, silhouette coefficient, and Calinski-Harabaz index.

The article is organized as follows: Section 2 reviews the related work on the ensemble clustering and gene expression data analysis. The theory of the proposed consensus function is presented in Section 3. The ensemble clustering algorithm is used to analyse the gene expression data and compared to some base clustering algorithm in section 4. The paper is concluded with some remarks and discussions.

2. Related Works

2.1. Clustering

As the one of the most important technologies of unsupervised learning, clustering deals with the data structure partition without given labels and is the basis for further learning. The main principles of clustering is presented in [11]:

- In the same cluster, instances must be similar as much as possible;
- Instances should be different as much as possible crossing clusters;
- Measurement for the similarity should be clear and explainable.

Traditional clustering algorithms can be divided into nine categories [11] and some of them are: partition clustering, hierarchical clustering, density-based clustering, fuzzy theory based clustering, and distribution-based clustering among them, partition clustering, hierarchical clustering and density-based clustering are more suitable for dealing with different kinds of data.

2.2. Ensemble Clustering

Strehl and Ghosh [7] formally proposed the definition of ensemble clustering in 2002. The main idea of ensemble clustering is to use multiple base clustering algorithms to group data samples and use one proper consensus function to integrate the multiple clustering results into the final result.

The core of an ensemble clustering algorithm is to design the consensus function that is responsible for integrating the clustering results generated by multiple base clustering algorithms. Existing consensus functions includes the co-association matrix algorithm [2], hypergraph partitioning algorithm [6], and finite mixture model algorithm [9].

Fred proposed a consensus function based on a co-association matrix [2]. It assigns two instances or samples to the same cluster if more than half base clustering algorithms group them into the same cluster. Each sample that is never assigned into any clusters in the above process will be put into a single-element cluster. Tang and Zhou proposed an ensemble method based on mutual information weighted voting [8]. The authors calculate the weights of base clustering learners by computing average mutual information with other clustering learners that means the degree of intimacy. Huang et.al. [4] consider the ensemble clustering problem as a binary linear programming problem and proposed a solution based on factor graph. Jiang [5] select five weight-calculation methods to implement a clustering ensemble algorithm with weighted voting. The results presented in the literature show that all clustering ensemble algorithms perform better than the base clustering algorithms.

The studies mentioned above reveal that an ensemble clustering algorithm has the advantages of improving clustering quality, robustness of solutions, knowledge reusing, and distributed computing.

2.3. Clustering on Gene Expression Data

Due to the wide variety of cells in the organism and the time-space specificity, gene expression data is more complex. At present, research on gene expression data mainly focuses on analysing the expression level of individual genes and gene combinations and inferring potential gene regulatory networks. Most gene combinations use the clustering analysis. In the field of genetic clustering, a lot of efforts have been spent.

Golub et al. [3] use self-organizing feature map neural networks and domain analysis methods to discover two types of human acute leukemia (acute myeloid leukemia and acute lymphoblastic leukemia) from the gene matrix. Wigle [10] et al. successfully differentiates the small cell lung cancer...
samples from the normal ones by the clustering algorithms and statistical analysis. Yuan et al. [13] discuss the clustering analysis of gene expression data and propose a large-scale gene expression data clustering algorithm based on the Isomap that applies the non-linear dimensionality reduction algorithm for isochronous feature mapping. Zhao et al. [15] employ the method of equal maximum subspace to cluster the coordinated genes.

In recent years, more and more people have used ensemble clustering approaches to study genetic related problems. In 2010, Xu et al. [12] use the ensemble clustering algorithm for gene expression data based on wavelet denoising and apply the micro-precision method to evaluate the new fuzzy ensemble clustering algorithm. An ensemble clustering algorithm framework is proposed by Zhang and Chen [14] to mine the intrinsic structure of noise-containing cancer gene datasets. Compared to the traditional base clustering algorithms, the experimental results show that the use of this ensemble clustering algorithm greatly improves the accuracy of cancer diagnosis.

3. Algorithm

3.1. Consensus Function

The labels are unknown in clustering problems because of the unsupervised learning properties. Nevertheless, the labels or resultant clusters produced by base clustering algorithms need to be matched and unified. The consensus function then comes to play, which integrates multiple clustering results yielded by base clustering algorithms to generate a final and unified clustering result.

In this paper, we combine the ideas of the co-association matrix and mutual information weighted voting, and proposes a new consensus function. The main idea is to calculate the co-association matrix of multiple clustering results. But it is different from Fred’s algorithm that believes all base clustering algorithms are equally important, we calculate the average mutual information of each base clustering algorithm to weigh the importance of these base clustering algorithms. The proposed consensus function is described as follows:

In the following discussions it is supposed \( k \) clusters to be built. Assume the \( i \)-th base clustering algorithm does clustering on the dataset having \( n \) samples and obtains the clustering result denoted by a labelled column vector \( R^i = (r_{i1}, r_{i2}, \ldots, r_{in})^T \). The value in the resultant vector represents the cluster IDs of the corresponding samples. If there are \( m \) base clustering algorithms, \( m \) labelled vectors of the clustering results will be generated. It then can be combined into an \( n \times m \) clustering result matrix \( R_{nm} \), represented as follows:

\[
R_{nm} = \begin{bmatrix}
    r_{11} & r_{12} & \cdots & r_{1m} \\
    r_{21} & r_{22} & \cdots & r_{2m} \\
    \vdots & \vdots & \ddots & \vdots \\
    r_{n1} & r_{n2} & \cdots & r_{nm}
\end{bmatrix}
\]  

As we mentioned earlier \( n \) is the number of samples in the dataset, \( m \) is the number of base clustering algorithms, and \( r_{ij} \) is the cluster ID between 0 and \( k-1 \) generated by the \( j \)-th base clustering algorithm on the \( i \)-th sample.

In Fred’s paper, the expression of co-association matrix \( C_{nm} \) is defined as follows:

\[
C_{nm} = \begin{bmatrix}
    c_{11} & c_{12} & \cdots & c_{1n} \\
    c_{21} & c_{22} & \cdots & c_{2n} \\
    \vdots & \vdots & \ddots & \vdots \\
    c_{m1} & c_{m2} & \cdots & c_{mn}
\end{bmatrix}
\]
where $C_{nn}$ is a symmetric matrix, the value of $c_{ij}$ stands for whether the $i$-th and $j$-th samples belong to the same cluster. 0 means they don't belong to the same cluster while 1 means the opposite. The formula to calculate the value of $c_{ij}$ is:

$$
c_{ij} = \begin{cases} 
1 & \sum_{k=1}^{n} s(r_{ik}, r_{jk}) > \left\lceil \frac{m}{2} \right\rceil \\
0 & \text{otherwise}
\end{cases}
$$

(3)

and

$$
s(x, y) = \begin{cases} 
1 & x = y \\
0 & \text{otherwise}
\end{cases}
$$

(4)

where $r_{ik}$, $r_{jk}$ are elements in the clustering result matrix $R_{nn}$.

According to the formula of the co-association matrix, we can conclude that all base clustering algorithms are equally important and their weights are the same, i.e., $1/m$. Nevertheless, the clustering ability of each base clustering algorithm is not exactly the same. Therefore, the mutual information is chosen here to measure the clustering effect and calculate the weight of each clustering algorithm.

Mutual information is an information measure method in the information theory. It measures the interdependence of two random variables. The usage of the mutual information in this paper is explained as follows:

The clustering result can be expressed with a vector whose length is $k$, i.e., $\{C_1, C_2, \ldots, C_k\}$, where element $C_i$ includes samples clustered into the $i$-th cluster. If there are $m$ base clustering algorithms, then $m$ such vectors will be generated. $\lambda^{(i)} = \{C_{i1}, C_{i2}, \ldots, C_{ik}\}$ and $\lambda^{(j)} = \{C_{j1}, C_{j2}, \ldots, C_{jk}\}$ are two vectors generated by the $i$-th and $j$-th base clustering algorithms, where element $C_{ik}$ includes $n_i$ samples, element $C_{jl}$ includes $n_j$ samples, and the number of same samples in both $C_{ik}$ and $C_{jl}$ is $n_{kl}$. If the total number of samples is $n$, then the mutual information of the $i$-th and $j$-th base clustering algorithms can be defined as follows:

$$
\phi_{ij}^{n} (\lambda^{(i)}, \lambda^{(j)}) = \frac{2}{n} \sum_{i=1}^{k} \sum_{j=1}^{k} n_{ij} \log \left( \frac{n_{ij}}{n_i n_j} \right)
$$

(5)

The average mutual information of the $i$-th base clustering algorithm is defined by (6):

$$
\beta_i = \frac{1}{m-1} \sum_{j=1,j \neq i}^{m} \phi_{ij}^{n} (\lambda^{(i)}, \lambda^{(j)}) (i = 1, 2, \ldots, m)
$$

(6)

When $\beta_i$ is larger, it means the $i$-th base clustering algorithm shares more common information with other base algorithms. One crucial part to build an effective ensemble learner is that the base learners should be diversified as much as possible [16]. To this end, we are going to deduct the weights in voting for the base clustering algorithms with larger $\beta_i$ values. By utilizing the average mutual information, we can define the weight of the $i$-th base clustering algorithm as follows:

$$
\omega_i = \frac{1}{\beta_i Z} (i = 1, 2, \ldots, m)
$$

(7)

where $Z$ is used to normalize the weights of all base algorithms to ensure the sum of all weights to be 1.

The weights of the base algorithms are applied to revise the calculation of the elements in a co-association matrix. Instead of using an equal weight for all base clustering learners, (7) is use to determine the weights and the value of $c_{ij}$ is defined by:
\[ c_\eta = \sum_{i=1}^{n} w_i s(R_i, R_a) \]  

At the end, we can construct the final clusters with the new co-association matrix. The procedure is: Pick the first row \([c_1, c_2, \ldots, c_m]\) of the co-association matrix \(C_{m \times m}\) and find all element \(c_i\) whose values are greater than 0.5 and label the corresponding samples as cluster “1”. Then get next row \([c_1, c_2, \ldots, c_m]\), find all elements \(c_i\) whose values are greater than 0.5 and do not belong to cluster “1”. The corresponding samples will be marked as cluster “2”. Continue this step until all samples are assigned into certain clusters and the final result is generated.

The above procedure might produce some conflicts: Assuming that \(s_a, s_b, s_c\) are three samples in the dataset. It could happen that \(s_a\) and \(s_b\) belong to the same cluster, while \(s_a\) and \(s_c\) belong to the same cluster. However, \(s_b\) and \(s_c\) do not belong to the same cluster. To resolve this issue, we compare the value of \(c_{ab}\) and \(c_{ac}\) in the co-association matrix. If \(c_{ab}\) is greater than \(c_{ac}\), \(s_a\) and \(s_b\) will be in the same cluster. Otherwise, \(s_a\) and \(s_c\) will be assigned to the same cluster.

The whole process of the ensemble clustering algorithm is summarized in Table 1.

### Table 1. The process of ensemble clustering.

| Input: N samples; M base clustering algorithms |
|---------------------------------------------|
| Output: Data clusters                       |
| Steps:                                       |
| 1. Initialized clustering result matrix \(R\), co-association matrix \(C\) and cluster index \(cn=1\); |
| 2. Produce data clustering results:          |
|   For \(i = 1\) to \(M\) do                |
|     2.1 Run the \(i\)-th base learner and produce a result vector \(R_i\); |
|     2.2 Update the result matrix \(R\) with vector \(R_i\); |
| 3. Calculate the weights of base clustering learners: |
|   For \(i = 1\) to \(M\) do:               |
|     Calculate average mutual information as the weight of the \(i\)-th learner using formulas (6) and (7). |
| 4. Produce the co-association matrix         |
|   For \(i = 1\) to \(N\) do:               |
|     For \(j = i\) to \(M\) do:             |
|       Calculate the element \(c_{ij}\) and \(c_{ji}\) in matrix \(C\) with formula (8); |
| 5 Generate the final clusters using co-association matrix \(C\) |
|   5.1 Pick the first row \(c_i\) in \(C\), find all elements \(c_i\) whose values are greater than 0.5 and label related samples as cluster \(cn\); |
|   5.2 Pick next row \(c_i\) that the \(i\)-th sample is not divided into any cluster, and find all elements whose values are greater than 0.5 and label related samples as clusters \(cn++\); |
|   5.3 Resolve any conflict using the solution introduced above; |
|   5.4 Repeat the step 5.2 and 5.3 until all elements are assigned into certain clusters. |
| 6. Return the final clustering result.        |

#### 3.2. Evaluation Metrics

To evaluate and compare the performances of the ensemble clustering and base clustering algorithms, different evaluation metrics are applied. For unlabelled datasets, the clustering result will be evaluated according to the principle that instances should be as similar as possible in the same cluster and different as much as possible across clusters. In this case, the metrics such as Silhouette Coefficient and Calinski-Harabaz Index will be employed. For labelled sample data, in addition to above two
metrics the Adjusted Rand Index and Adjusted Mutual Information are selected to evaluate clustering results. For all above selected metrics, the higher the values are, the better the clustering performance.

4. Experiments

4.1. Data and Data Preprocess
The gene expression datasets used in our experiments are downloaded from the public repository that have been extensively studied including the GEMS (Gene Expression Model Selector) and GEO (Gene Expression Omnibus) databases. Since these datasets have been extensively studied, we can relatively easily evaluate the performance of our approach. The sample size, attribute dimensions, and number of clusters of all datasets are described in table 2 and 3:

| Dataset   | Size | Dimension | #Clusters |
|-----------|------|-----------|-----------|
| Tumors    | 308  | 15009     | 26        |
| BrainTumor| 50   | 10367     | 4         |
| Leukemia  | 72   | 11225     | 3         |
| Lung_cancer| 203 | 12600     | 5         |
| SRBCT     | 83   | 2308      | 4         |
| Prostate Tumor | 102 | 10509    | 2         |
| DLBCL     | 77   | 5469      | 2         |

Table 3. GEO dataset

| Dataset  | Size | Dimension | #Clusters |
|----------|------|-----------|-----------|
| GDS3268  | 202  | 44290     | 2         |
| GDS3795  | 200  | 54675     | 2         |
| GDS4456  | 93   | 54675     | 2         |
| GDS6247  | 40   | 45281     | 2         |

Having observed the data, both duplicate and missing values are found in the GEO dataset. The duplicated records are removed and only one record is retained. For missing values, different preprocessing methods are applied according to the missing severity: if more than 50% of the samples miss one attribute value, then this attribute will be removed. If less than 20% of the samples miss an attribute value, then it will be filled with the average. For the case in between, the corresponding attribute value will be filled by applying k-NN algorithm.

Considering that most clustering algorithms are disturbed by outliers, for instance, K-Means algorithm, we deliberately select K-Means as the base clustering algorithm of the ensemble clustering algorithm in this experiment. In such way we can identify if the ensemble clustering algorithm is indeed able to handle the outliers more effectively. In the experiment, based on the analysis of the distribution of box plots for each dataset, any sample whose attribute values over 99 percentiles or below 1 percentiles is defined as an outlier. Further, we utilize the stochastic property of K-Means to run several K-Means in parallel on the same dataset. In this case we are able to simulate multiple base clustering algorithms and all above discussed methodologies will be applicable.

Before the experiment is conducted, the samples of the dataset is normalized so that the distribution of the corresponding dataset can be adjusted to a standard Gaussian distribution to overcome the issue of wide fluctuations in attribute values that could impact the underlying algorithms negatively.

4.2. Computational Experiments
As mentioned above the base clustering learner used in the ensemble clustering algorithm is K-Means algorithm that is running several parallel instances on the same dataset. Our proposed approach is benchmarked against some well-known base clustering algorithms including K-Means, DBSCAN, and
hierarchical clustering algorithms. Due to the poor performance of DBSCAN on the datasets even after careful parameter tuning, we decided to drop it from the further experiments. The following analysis will not contain the results obtained by DBSCAN.

Since all samples in GEMS dataset are labelled, the Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), Silhouette Coefficient (S), and Calinski-Harabaz Index (CH) are applied to evaluate the clustering results. The Figure 1 and Figure 2 are the histograms of different metrics for the performances of the ensemble clustering and base clustering algorithms on GEMS dataset.

**Figure 1.** Adjusted Mutual Information and Adjusted Rand Index on GEMS datasets

**Figure 2.** Silhouette Coefficient and Calinski-Harabaz Index on GEMS datasets

In the diagrams the horizontal axis represents different datasets while the vertical axis represents the metrics values. We use different colours to represent the metric of different algorithms, specifically, blue, orange, and green are for K-Means, hierarchical, and ensemble clustering algorithms respectively.

The histograms demonstrate that the ensemble clustering algorithm outperforms two base clustering algorithms. As we point out in section 3, the larger the value of a metric is, the better the performance. And the result of K-Means is better than those obtained by the hierarchical clustering. The average values of the metrics for all clustering algorithms on GEMS dataset are presented in table 4:

**Table 4.** Average metric value in each dataset

| Algorithm            | ARI | AMI  | S   | CH   |
|----------------------|-----|------|-----|------|
| KMeans               | 0.205 | 0.219 | 0.219 | 28.35 |
| Hierarchical clustering | 0.185 | 0.095 | 0.095 | 8.952 |
| Ensemble clustering  | 0.258 | 0.316 | 0.29 | 30.209 |

Again, it can also be seen that the ensemble clustering performs on GEMS dataset overwhelmingly.

The performances of three clustering algorithms are also compared using GEO dataset. Since the data is not labelled, we need to use two metrics, namely, Silhouette Coefficient and Calinski-Harabaz Index, to evaluate the performance. The results are shown in the Figure 3, the representations of axes and the colours are exactly the same as those in figures 1 and 2.
From Figure 3, it is not hard to recognize that the ensemble clustering algorithm performs the best following by K-Means, and the hierarchical clustering algorithm is the worst.

Based on the experimental results on GEMS and GEO datasets, we can conclude that both hierarchical clustering and DBSCAN clustering algorithms are not effective for the underlying gene expression data. The performance of K-Means is the best among the base clustering algorithms. The ensemble clustering algorithm proposed in this paper has the best performance. According to our observations, the results obtained by the ensemble clustering algorithm are more stable.

5. Conclusion
This paper focuses on developing an ensemble clustering algorithm applicable for handling the clustering gene expression datasets. Based on mutual information and co-association matrix, we develop a weighted voting mechanism to implement the ensemble clustering algorithm. With the mutual information, different base clustering algorithms are assigned with the different weights during the generation of final results. The weights represent the importance of associated algorithms.

The gene expression datasets (GEMS and GEO) are utilized to conduct the computational experiments. The base clustering algorithms including K-Means and hierarchical clustering are used to benchmark against the proposed ensemble clustering algorithm. Various metrics are employed to evaluate the performances of the benchmarked clustering algorithms. The experiments validate the superiority of the proposed ensemble clustering algorithm in terms of solution quality and robustness.

The current ensemble clustering algorithm might be lack of outstanding diversity. In the future research we are looking into using baser clustering algorithms to generate an ensemble clustering algorithm that could be able to produce better solutions and possess strong generalization capability.

6. Acknowledgments
This research is supported partially by China Intelligent Urbanization Co-Creation Center for High Density Region and Tongji Architectural Design (Group) Co., Ltd. under grant No. CIUC20150011.

7. References
[1] Farina L, Santis A D, Salvucci S, Morelli G and Ruberti I 2008 Embedding mRNA stability in correlation analysis of time-series gene expression data J Plos Computational Biology 4 e1000141.
[2] Fred A 2001 finding consistent clusters in data partitions Multiple Classifier Systems 309-18.
[3] Golub T R, Slonim D K and Tamayo P 1999 Molecular classification of cancer: class discovery and class prediction by gene expression monitoring Science 286 531-37.
[4] Huang D, Lai J and Wang C D 2016 Ensemble clustering using factor graph Pattern Recognition 50 131-42.
[5] Jiang Z, Hou Y, Min and W U 2018 Clustering ensemble with weighted voting based on feature correlation Computer Engineering & Applications.
[6] Lotfifar F and Johnson M A 2015 Multi–level hypergraph partitioning algorithm using rough set clustering European Conference on Parallel Processing 159-70.
[7] Strehl A and Ghosh J 2002 Cluster ensembles - a knowledge reuse framework for combining multiple partitions Journal of Machine Learning Research 3 583- 617.
[8] Tang W and Zhou Z 2005 Bagging-based selective clusterer ensemble *Journal of Software* 496-502.

[9] Topchy A P, Jain A K and Punch W F 2004 A mixture model for clustering ensembles. *Siam International Conference on Data Mining*.

[10] Wigle D A, Jurisica I and Radulovich N 2002 Molecular profiling of non-small cell lung cancer and correlation with disease-free survival *Cancer Res* 62 3005-8.

[11] Xu D and Tian Y 2015 A comprehensive survey of clustering algorithms *Ann Data Science* 2 165–93.

[12] Xu Y, Peng B and Deng Y 2010 Approach of clustering ensemble in gene expression data based on wavelet de-noising *Journal of Chongqing Medical University* 35 718 -720.

[13] Yuan Y, Ji X and Sun Z 2004 Application of isomap for cluster analyses of gene expression data *Journal of Tsinghua Univ (Sci. &Tech)* 44 1286-9.

[14] Zhang X and Chen H 2015 Cluster ensemble algorithm based on dual neural gas applied to cancer gene expression profiles *Journal of Biomedical Engineering* 32 93 - 8.

[15] Zhao Y, Xu J, Yu W, Chen L, Wang B and Yu G 2008 Maximal subspace coregulated gene clustering *IEEE Transactions on Knowledge and Data Engineering* 20.

[16] Zhou Z H 2015 *Ensemble learning* Encyclopedia of biometrics 411- 41.