Feature Filtering Spectral Clustering Method Based on High Dimensional Online Clustering Method

Zizhou Feng[1], Yujian Gu[1], Bin Yang[2], Baitong Chen[3], and Wenzheng Bao[1(✉)]

[1] School of Information Engineering, Xuzhou University of Technology, Xuzhou 221000, China
[2] School of Information Science and Engineering, Zaozhuang University, Zaozhuang 277160, China
[3] Xuzhou No. 1 People's Hospital, Xuzhou 221000, China

Abstract. Golgi is an important eukaryotic organelle. Golgi plays a key role in protein synthesis in eukaryotic cells, and its dysfunction will lead to various genetic and neurodegenerative diseases. In order to better develop drugs to treat diseases, one of the key problems is to identify the protein category of Golgi apparatus. In the past, the physical and chemical properties of Golgi proteins have often been used as feature extraction methods, but more accurate sub-Golgi protein identification is still challenged by existing methods. In this paper, we use the tape-bert model to extract the features of Golgi body. To create a balanced dataset from an unbalanced Golgi dataset, we used the SMOTE oversampling method. In addition, we screened out the important eigenvalues of 300 dimensions to identify the types of Golgi proteins. In 10-fold cross validation and independent test set test, the accuracy rate reached 90.6% and 95.31%.

Keywords: Golgi appratus · Malonylation · SMOTE · Protein

1 Introduction

In recent years, spectral clustering has become one of the most popular clustering algorithms [1]. It is easy to implement, can be solved effectively by standard linear algebra software, and is often better than traditional clustering algorithms, such as k-means algorithm [2].

Common spectral clustering algorithms usually include loading data, calculating Euclidean distance to obtain distance matrix, calculating adjacency matrix W and degree matrix D through distance matrix, so as to obtain Laplacian matrix $L = D - W$, then decomposing Laplacian matrix L to obtain characteristic matrix, and then clustering with k-means algorithm to obtain clustering results [3–6]. This spectral clustering algorithm is easy to understand and implement, but it has the disadvantages of slow running speed and low precision, and there is still a lot of optimization space [7–9].

There is a spectral clustering method for high-dimensional online clustering [10–12]. By further optimizing the Laplacian matrix (i.e. feature matrix), the feature matrix is

D. Jiang and H. Song (Eds.): SIMUtools 2021, LNICST 424, pp. 157–164, 2022.
https://doi.org/10.1007/978-3-030-97124-3_14

processed by using cropdiagonal, Gaussian blur, rowwise threshold, symmetry, diffusion and rowwise normalize, The feature matrix similar to Laplacian matrix is obtained, which can provide real-time and effective clustering for data [13–15].

The purpose of this paper is to improve the speed and accuracy of clustering by feature processing on the basis of the high-dimensional online clustering and removing the features with low correlation coefficient through the correlation coefficient matrix.

2 Methods and Materials

2.1 Optimization Idea of Spectral Clustering Model

Spectral clustering is a clustering algorithm based on graph theory. Therefore, the standard spectral clustering algorithm first regards the data as a graph. If the data is two-dimensional, it can be intuitively represented by image. If the data is multi-dimensional or even high-dimensional, it can only be represented by abstract formula.

If the intra cluster similarity is high and the inter cluster similarity is low, the clustering performance is better. Therefore, the standard of optimizing the clustering model is to improve the cluster similarity and reduce the inter cluster similarity.

The optimization idea of spectral clustering model is also based on this standard.

The optimization method of spectral clustering model is to minimize the objective function.

2.2 The Definition of Graph and Adjacency Matrix and Degree Matrix

Graph G is composed of the set of points V (vertex) and the set of edges e (edge), that is, g = (V, e), where V is the data set V = {V1, V2, ... VN}, e is the weight of the sample point VI and the sample point VJ, represented by Wij, Wij equal to 0 means that the sample point VI is not connected with the sample point VJ. Therefore, the directed adjacency matrix w of the graph for the data set with capacity n is expressed as: $W = \left(w_{ij}\right)_{i,j=1,\ldots,n}$ Weight of undirected graph $w_{ij} = w_{ji}$.

The undirected weight W in the figure above is expressed as:

$$\begin{pmatrix} w_{11} & w_{12} & 0 & w_{14} & 0 \\ w_{21} & w_{22} & w_{23} & 0 & 0 \\ 0 & w_{32} & w_{33} & 0 & w_{35} \\ w_{41} & 0 & 0 & w_{44} & w_{45} \\ 0 & 0 & w_{53} & w_{54} & w_{55} \end{pmatrix} \tag{1}$$

definition d_i It is the sample point v_i Degree of freedom:

$$d_i = \sum_{j=1}^{n} w_{ij} \tag{2}$$

The meaning of sample point degree is the sum of all weights connected with the sample point.

The degree of all sample points in the dataset is defined as the degree matrix D:

$$D = \begin{pmatrix} d_1 & & & \\ & d_2 & & \\ & & \ddots & \\ & & & d_n \end{pmatrix} \tag{3}$$

The matrix D is a diagonal matrix and the off diagonal elements are all zero.

2.3 Representation of Adjacency Matrix

The weight of adjacency matrix is the similarity between samplesεIn this paper, Euclidean distance is used to express the similarity between sample points

$$d(x, y) := \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \tag{4}$$

2.4 Laplacian Matrix and Its Properties

Laplacian matrix L is the basis of spectral clustering algorithm. There are two kinds of Laplacian matrices and their attributes, namely non standardized Laplacian matrix and standardized Laplacian matrix.

1 Nonstandardized Laplacian matrix.
The non Laplacian matrix is defined as the difference between the degree matrix D and the adjacency matrix W. the expression is as follows:

$$\mathrm{L} = \mathrm{D} - \mathrm{W} \tag{5}$$

We have two methods to define the standardized Laplacian matrix L_{sym} and L_{rw}, defined as:

$$\begin{aligned} L_{sym} &= D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \\ L_{rw} &= D^{-1} L = I - D^{-1} W \end{aligned} \tag{6}$$

2.5 The Meaning of Cut Graph of Undirected Graph

An undirected graph is composed of sample points and edges. The clustering of data set can be regarded as the segmentation of undirected graph. Suppose that graph G contains two connected subsets A and B after segmentation, then the weight of tangent graph between AB is as follows:

$$cut(A, B) = \sum_{i \in A, j \in B} w_{ij} \tag{7}$$

among Denotes the adjacency matrix of graph G.

If G is cut into k connected subsets AI(I = 1, 2, . . . , K), the simplest method is to minimize the following formula:

$$\mathrm{cut}(A_1, \ldots, A_k) := \sum_{i=1}^{k} \mathrm{cut}\left(A_i, \overline{A}_i\right) \tag{8}$$

among $\overline{A}_i$ express A_i The complement of.

This segmentation method only considers minimizing the similarity between clusters, and does not consider the similarity within clusters, so this segmentation standard is not accurate, so we need to optimize the segmentation method, and there are two kinds of optimal segmentation methods: ratiocut segmentation and ncutt segmentation. Ncutt cut graph is also called standardized spectral clustering algorithm, and ratiocut cut graph is called non standardized spectral clustering algorithm.

$$RatioCut(A_1, A_2, \ldots, A_k) = \sum_{i=1}^{k} \frac{cut\left(A_i, \overline{A_i}\right)}{|A_i|} \tag{9}$$

$$Ncut(A_1, \ldots, A_k) = \sum_{i=1}^{k} \frac{cut\left(A_i, \overline{A_i}\right)}{vol(A_i)} \tag{10}$$

2.6 The Choice of Laplacian Matrix

There are two kinds of algorithms for Laplacian matrix. Which algorithm to choose is a basic problem of spectral clustering. If the graph is regular and the degrees of most sample points are approximately equal, it is feasible to choose any kind of Laplacian matrix. If the degree difference of most sample points is large, it is recommended to use the standardized Laplacian matrix. Because the nonstandardized Laplacian matrix corresponds to the ratiocut cut graph, the similarity within the cluster described by the ratiocut cut graph is the number of samples contained in the cluster |a|, the standardized Laplacian matrix corresponds to the ncutt cut graph, and the ncutt cut graph describes the similarity within the cluster as Vol (a). Because Vol (a) is better than |a| in reflecting the similarity within clusters, this paper chooses the standardized Laplacian matrix.

2.7 Selection of the Number of Cluster Classes

The first problem of spectral clustering algorithm is the selection of the number of clusters. The common way is to use the heuristic eigenvalue difference search (eigengap heuristic), meaning: if the first k eigenvalues are very small, and the K + 1 eigenvalue is quite different from the previous eigenvalue, then the number of clusters is K. Let G be partitioned into k connected subsets without intersection, then K eigenvalues are equal to 0 and K + 1 eigenvalues are greater than 0. Therefore, we can assume that the smaller the eigenvalue is, the better the clustering performance is, and select the number of clusters with very small eigenvalue as the number of clusters. In this paper, the high-dimensional online clustering algorithm based on links selects cluster K automatically according to the samples.

2.8 Feature Filtering

For a specific learning algorithm, which feature is effective is unknown. Therefore, it is necessary to select the useful features from all the features. And in practical application, the problem of dimension disaster often appears. If only some of the features are selected to build the model, the running time of the learning algorithm can be greatly reduced, and the interpretability of the model can be increased.

The principle of feature selection is to obtain as small a feature subset as possible, not to significantly reduce the classification accuracy, not to affect the classification distribution, and the feature subset should be stable and adaptable.

There are many methods of feature selection, such as chi square test, information gain, correlation coefficient and so on. This paper adopts the correlation coefficient method to judge the correlation coefficient and a certain threshold between each column of data (each column of data is a different sample value represented by the same feature) and the label, It can be considered that the column features have little correlation with the results, which will affect the classification effect, and the column is removed, that is to complete a feature filtering.

Feature filtering has the following advantages: first, reduce the number of features, dimension reduction; second, reduce the difficulty of learning tasks, improve the efficiency of the model; third, make the model more pan Chinese ability, reduce over fitting; fourth, enhance the understanding between features and eigenvalues.

3 Algorithm

3.1 Standard Spectral Clustering Algorithm

Spectral clustering is a clustering algorithm based on graph theory. Therefore, the standard spectral clustering algorithm first regards the data as a graph. If the data is two-dimensional, it can be intuitively represented by image. If the data is multi-dimensional or even high-dimensional, it can only be represented by abstract formula.

After the data is regarded as undirected weight graph, the specific process of spectral clustering is as follows:

1. By Euclidean distance or ε-Neighborhood method, k-nearest neighbor method and other methods are used to calculate the distance between each node and get the distance matrix.
2. The adjacency matrix A and degree matrix D are calculated by the distance matrix.
3. The nonstandardized Laplacian matrix $L = D - A$ is obtained.
4. Normalized Laplacian matrix: $l \rightarrow D - 1 / 2ld - 1 / 2$.
5. The eigenvector HN is obtained by eigendecomposition of the normalized Laplacian matrix.
6. The feature vector HN is sent to kmeans clustering as a sample.
7. The clustering result $c = (C1, C2, \ldots, CN)$ is obtained.

Spectral clustering is a kind of clustering method based on data similarity matrix. It defines the optimization objective function of subgraph partition, introduces indicator

variables, and transforms the partition problem into solving the optimal indicator variable matrix HH.Then, by using the properties of Rayleigh entropy, the problem is further transformed into solving the K minimum eigenvalues of Laplacian matrix. Finally, as some expression of samples, the traditional clustering method is used for clustering.

3.2 High Dimensional Online Clustering Method

The spectral clustering method in this paper is based on this method. This spectral clustering method is called links, which aims to cluster the unit vectors of high-dimensional Euclidean space online. This algorithm is suitable for the situation that the data need to be effectively clustered when the data stream enters. What this paper focuses on is the excellent running speed and accuracy of this method when processing high-dimensional data.

This method uses six default optimization methods: cropdiagonal, Gaussian blur, rowwise threshold, symmetry, diffuse and rowwise normalize to refine the feature matrix, so as to get more accurate results.The specific steps are as follows:

1. The similarity matrix affinity is calculated by sample data
2. Six default optimization methods are used to optimize the similarity matrix affinity
3. The similarity matrix affinity is decomposed into feature matrix and feature vector
4. Through the characteristic matrix, the characteristic vector, the maximum number of clusters and the minimum number of clusters, the number of clusters K is obtained.
5. The first k minimum eigenvalues of feature vector are taken and sent to kmeans clustering
6. The result c = (C1, C2, . . . , CN).

3.3 High Dimensional Spectral Clustering Algorithm Based on Feature Filtering

In this paper, based on the links spectral clustering algorithm, the feature selection function is added. Many processed data are high-dimensional data, some data samples may have more than ten, dozens or hundreds of thousands of dimensional data features, some of which have little correlation with the clustering results, that is to say, it will play a role of interference, resulting in the accuracy of clustering results. The innovation of this paper lies in the feature selection of data, filtering out the data with low correlation of clustering results, so as to improve the accuracy of clustering results.

The algorithm is as follows

1. Extract and separate the data information and tags contained in the data
2. Obtain the correlation coefficient (COR) between each column of data (each column of data is a different sample value represented by the same feature) and the tag
3. The average exp of all correlation coefficients is calculated as the threshold to judge the correlation
4. Traverse the correlation coefficient of each column of data, when corn $<$ exp, delete the column, otherwise keep the column
5. Get the filtered new data
6. The new data is sent to the links spectral clustering method for clustering
7. The result c = (C1, C2, . . . , CN).

4 Conclusion

In this paper, three spectral clustering methods, standard spectral clustering, high-dimensional online clustering (links) and feature filtered high-dimensional spectral clustering, are compared in terms of algorithm accuracy and operation time, It can be seen from the above figures and that the precision of the feature filtered high-dimensional spectral clustering is higher than the high-dimensional online clustering (links) and table quasi spectral clustering, and the clustering speed is much faster than the standard spectral clustering algorithm. It can be seen that this model not only retains the high speed of high-dimensional online clustering (links), but also improves the accuracy of clustering results (Table 1).

Table 1. The comparison of three method

Method	Standard spectral clustering	High dimensional online clustering	High dimensional spectral clustering based on feature filtering
Run time	2503.372078180313 s	769.8853192329407 s	775.8828499317169 s

Acknowledgement. This work is supported by the fundamental Research Funds for the Central Universities, 2020QN89, Xuzhou science and technology plan project (KC19142), the talent project of 'Qingtan scholar' of Zaozhuang University, Jiangsu Provincial Natural Science Foundation, China (SBK2019040953), Youth Innovation Teamof Scientific Research Foundation of the Higher Education Institutions of Shandong Province, China (2019KJM006), the Key Research Program of the Science Foundation of Shandong Province (ZR2020KE001), the PhD research startup foundation of Zaozhuang University (2014BS13) and Zaozhuang University Foundation (2015YY02), the Natural Science Foundation of China (61902337), Natural Science Fund for Colleges and Universities in Jiangsu Province (19KJB520016), Xuzhou Natural Science Foundation KC21047 and Young talents of science and technology in Jiangsu.

References

1. Molinie, B., Giallourakis, C.C.: Genome-wide location analyses of N6-Methyladenosine modifications (m(6)A-Seq). Methods Mol. Biol. **1562**, 45–53 (2017)
2. Nye, T.M., van Gijtenbeek, L.A., Stevens, A.G., et al.: Methyltransferase DnmA is responsible for genome-wide N6-methyladenosine modifications at non-palindromic recognition sites in Bacillus subtilis. Nucleic Acids Res. **48**, 5332–5348 (2020)
3. O'Brown, Z.K., Greer, E.L.: N6-methyladenine: a conserved and dynamic DNA mark. In: Jeltsch, A., Jurkowska, R. (eds.) DNA Methyltransferases-Role and Function, vol. 945, pp. 213–246. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-43624-1_10
4. Zhang, G., et al.: N6-methyladenine dna modification in drosophila. Cell **161**(4), 893–906 (2015)
5. Janulaitis, A., et al.: Cytosine modification in DNA by BCNI methylase yields N4-methylcytosine. FEBS Lett. **161**, 131–134 (1983)

6. Unger, G., Venner, H.: Remarks on minor bases in spermatic desoxyribonucleic acid. Hoppe-Seylers Z. Physiol. Chem. **344**, 280–283 (1966)
7. Fu, Y., et al.: N6-methyldeoxyadenosine marks active transcription start sites in Chlamydomonas. Cell **161**, 879–892 (2015)
8. Greer, E.L., et al.: DNA methylation on N6-adenine in C. elegans. Cell **161**, 868–878 (2015)
9. Zhang, G., et al.: N6-methyladenine DNA modification in Drosophila. Cell **161**, 893–906 (2015)
10. Wu, T.P., et al.: DNA methylation on N6-adenine in mammalian embryonic stem cells. Nature **532**, 329–333 (2016)
11. Xiao, C.L., et al.: N-methyladenine DNA modification in the human genome. Mol. Cell **71**, 306–318 (2018)
12. Zhou, C., et al.: Identification and analysis of adenine N6-methylation sites in the rice genome. Nat. Plants **4**, 554–563 (2018)
13. Chen, W., et al.: i6mA-Pred: identifying DNA N6-methyladenine sites in the rice genome. Bioinformatics **35**, 2796–2800 (2019)
14. Almagor, H.A.: A Markov analysis of DNA sequences. J. Theor. Biol. **104**, 633–645 (1983)
15. Borodovsky, M., et al.: Detection of new genes in a bacterial genome using Markov models for three gene classes. Nucleic Acids Res. **17**, 3554–3562 (1995)