2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Some New Attempts to Process Biological Data

Shuxun Yang, Mingpu Li, Jun Luo, Yupeng Lu, Chao Yan, Xu-Qing Tang*
School of Science
Jiangnan University
Wuxi 214122, Jiangsu, China
Email: txq5139@jiangnan.edu.cn

***Abstract*—The purpose of this paper is to realize system analysis and algorithm design for biological data. In this paper, primary bladder cancer is taken as a typical example, the structure of the system is extracted by hierarchical clustering method, and the function of the system is mined by convolutional neural network technology. Based on these methods, a complex system structure analysis model and an algorithm are constructed to study the big data system. Furthermore, the feasibility study of relevant theories and methods are carried out while the application and expand of technology are mentioned, combined with the actual data. The effectiveness and practicability of the algorithm and system are also verified by simulation.**

Keywords-Hierarchical Clustering; Convolutional Neural Network; Primary Bladder Cancer; Feature Extraction; Function Mining; Fuzzy Clustering

I. INTRODUCTION

In the information age, big data plays an increasingly prominent role in scientific and technological progress. In the field of biology, big data has been existed for a long time. At the beginning of the 21st century, the completion of the Human Genome Project marked the beginning of the era of life sciences, which meant a new paradigm of life science research emerged. Nowadays, biological resources have become the strategic resources of various countries, like the biological resources themselves, the biological information has become an important strategic resource as well. In the era of big data, this project adopts research methods such as hierarchical clustering and convolutional neural network to carry out the study on the systematic analysis and algorithm design of biological data.

Clustering is the process of grouping a set of data objects into multiple groups or clusters so that objects within a cluster have high similarity. Hierarchical clustering, is an effective way to extract dynamic structure information of complex systems, which is widely used in the area of biology, agriculture, finance, medical and so on. There are many studies based on hierarchical clustering. For example, Eisen [1] used hierarchical clustering for co-expression analysis of genes for the first time and then used it in the study of yeast gene co-expression. At the same time, many algorithms improved based on hierarchical clustering are also proposed. The BIRCH algorithm dynamically builds a cluster feature tree by constantly inserting the least distant object. CURE algorithm [2] uses the representative point to merge with its nearest object, avoids the spherical expansion of the center point and radius, and strengthens the control of the isolated point. Xu-Qing Tang [3] and Ping Zhu in Jiangnan University developed some hierarchical clustering problems and analysis for fuzzy proximity relation by using rigorous mathematical descriptions based on granular space. Besides, the K-Modes algorithm, extended by the K-Means algorithm, uses attribute dissimilarity instead of numeric distance; the ROCK algorithm uses the similarity function SIM and the interconnectivity between the data to arrive at a data similarity matrix; and the MPM algorithm uses the similarity function defined by the cosine similarity measurement to give the data similarity matrix. These ideas solved the problem that the hierarchical clustering algorithm can't deal with the classified attribute data effectively, and made the hierarchical cluster more rigorous and widely applicable. These studies provided the foundation of theoretical research and algorithm design for the dynamic structure information extraction of this project.

The research achievements of Convolutional Neural Network(CNN) in biology, natural language, computer technology and so on are also distinguished. After entering this century, with the introduction of deep learning theory and the improvement of numerical computing equipment, convolutional neural network has been rapidly developed and widely used in computer vision, natural language processing and other fields. In 2017, Müller [4] et al. of Swiss Federal Institute of Technology Zurich developed Generative Recurrent Neural Network(RNN) to generate new peptide chains through LSTM-based generative model training. It could be used to facilitate the construction of peptide chain library. In 2017, Klukowski of University of Wroclaw, Poland, and others used a six-layer CNN model to identify protein's Nuclear Magnetic Resonance(NMR) data's peak value, and predicted the peptide chain corresponding to the spectra. Tran [5] et al. of the University of Waterloo in Canada developed the DeepNovo model in conjunction with the CNN model and the RNN model, using protein mass spectrometry data to build a model to sequence peptide chain from scratch. The model achieved almost 100% accuracy in protein sequence prediction. Three-generation sequencing technology is the main future direction. Yang [6] et al. used the GRU-based two-way RNN model (GRU-BRNN) to identify the enhancers of non-coding function in the DNA sequence, which obtained good results. In fact, these studies provide substantial supports of theory and algorithm design for the structural-based functional mining of this project.

On the basis of the existed research, this project uses hierarchical clustering to extract the structure of the system, and then the convolutional neural network is used to mine the function of the system. These are to construct the complex system's structure analysis model and algorithm

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00084

design, and to develop big data system. Combined with the actual data, the feasibility and comparative study of the relevant theories and methods are carried out.

The purpose of our study is to develop more effective biological data system analysis method and algorithm design by using hierarchical clustering and convolutional neural network techniques, and to provide more effective study methods and approaches for biological data processing. This paper is a summary of our recent studies, and the specific experimental steps and procedures will be elaborated in the following papers. This paper is organized as follows. In section 2, we selected primary bladder cancer as the research object and pretreated it, and then applied multiple hierarchical clustering methods to the differential expression matrix of primary bladder cancer to extract feature information. In section 3, according to the feature information extracted above, the convolutional neural network is used to mine the function of the system. In section 4, the application and expand of hierarchical clustering and convolutional neural network are discussed. In section 5, the conclusions of this paper are given.

II. HIERARCHICAL CLUSTERING

In order to process biological data better with hierarchical clustering and convolution neural network technology, we compared various biological data, such as colon cancer data set, breast cancer data set, acute leukemia data set and so on, and then selected data set called GSE13507, that is, primary bladder cancer data set, from the National Biotechnology Information Center of the United States finally. The data set includes 165 primary bladder cancer samples, 23 recurrent non-muscle invasive tumor tissues, 58 normal looking bladder mucosae surrounding cancer and 10 normal bladder mucosae. To improve the quality of the experiment, we preprocess the dataset as follows.

Firstly, for the data in the expression matrix, the data of one probe corresponding to multiple genes and the data of probe without corresponding genes are removed. Secondly, the data of multiple probes corresponding to one gene are averaged. Then, we carried out the quantile standardization. According to the types of cells, they were divided into four groups, and the experimental design matrix of bicolor microarray was constructed according to the RNA target information. Furthermore, we analyzed differential expression and established the linear model of expression matrix. On this basis, the empirical Bayes analysis of differential expression is carried out while the high expression gene fitted by the linear model is found. Finally, the above results are sorted according to logarithm of fold change, and the final difference expression matrix is obtained. The difference expression matrix eliminates the redundant part of the original data set, and it can also describe the characteristics of the data set more effectively and concisely.

According to the differential expression matrix of primary bladder cancer obtained by the above pretreatment, we can extract the dynamic structure information of the system by hierarchical clustering. In the first place, the stratification boundary of biological data was determined. According to the characteristics of biological data, the stratification of data can be studied. And then, stratified analysis was performed to analyze the strength of association with cell function at different levels, so as to effectively distinguish and control the confounding bias, and extract the structural information of biological data.

A. Fuzzy Clustering

At first, we use fuzzy clustering analysis to extract information. The final dynamic clustering graph can be obtained by transitive closure and λ truncated set. The clustering process can be expressed as follows. First, data standardization is carried out, such as translation standard deviation transformation, translation range transformation, logarithm transformation. Second, calibration is started, that is, to establish fuzzy similarity matrix, which can be divided into similarity coefficient method, distance method and subjective scoring method. As is known to us, the similarity coefficient method includes the number product method, angle cosine method, correlation coefficient method, exponential similarity coefficient method, maximum minimum method, arithmetic average minimum method, geometric average minimum method, and so on. And the distance method includes direct distance method, reciprocal distance method, and exponential distance method. Third, the flat method is used to find transitive closure. Fourth, λ truncated sets can be dynamically clustered. According to the clustering results, we can draw the dynamic clustering graph. Based on the above process, we could write the code to construct the fuzzy similarity matrix by selecting the maxima and minima method to solve the transitive closure, and use the value of λ to cluster the differential expression matrix for primary bladder cancer.

B. Fuzzy C-Means Clustering

Fuzzy clustering analysis based on objective function is called fuzzy c-means clustering method, abbreviated as FCM. FCM algorithm belongs to partition clustering algorithm, which uses fuzzy method to deal with clustering problem. It starts from an initial partition, and needs to specify the number of clusters in advance. It also needs to define an optimal clustering standard, namely the objective function, as a cost function to measure the distribution of various samples. FCM divides N data vectors into C fuzzy classes, which are represented by the clustering center of each class. Through repeated iterative operation, the error value of the objective function is gradually reduced. When the objective function value converges, the final clustering result will be obtained. Based on the formula of FCM, we wrote programs for FCM clustering of the differential expression matrix of primary bladder cancer. According to the dynamic clustering diagram, we chose 100 as the cluster number of c-means clustering, and 464 data sets of the differential expression matrix of primary bladder cancer could be divided into 100 categories by comparing their characteristics. There are many ways to extract the information features of system structure. The clustering center can be used to represent the feature information as well as the data vector nearest to the clustering center. In this section, we use the clustering center to represent the extracted system structure information and the complete feature information will be given in the following paper.

C. Hierarchical Clustering

Hierarchical clustering is a very intuitive algorithm. As the name implies, clustering is carried out layer by layer. Small clusters can be merged and clustered from bottom to top, and large ones can also be divided from top to bottom.

The process of hierarchical clustering can be represented as follows. First of all, the similarity between the data is determined, that is, each data point in the data set is determined, and then the distance between a representational object is defined to represent the similarity between the data. For example, the Euclidean distance is the most often used in the clustering of points on the simplest plane. Secondly, after determining the degree of difference between objects, namely, the distance, we can make use of the linkage function of MATLAB, or the self-programmed program, to generate hierarchical clustering tree. The vertical axis height in hierarchical clustering represents the distance between two child nodes in the current clustering, and the horizontal axis marks the subindexes of each data point. Finally, some functions, such as cluster, clusterdata, cophenet, inconsistent and so on, provided by MATLAB, are used to test the consistency or difference between the clustering tree generated under a certain algorithm and the actual situation for optimization.

According to the clustering results obtained, the 464 data sets of the differential expression matrix of primary bladder cancer were divided into 100 categories by their features. Then, the data vectors closest to the cluster center were searched as the feature information of the system, and 100 characteristic genes identified for primary bladder cancer could be obtained, which will be shown completely in the following paper we expanded.

III. Convolutional Neural Network

As one of the representative algorithms of deep learning, convolutional neural network is a kind of feedforward neural network with deep structure including convolution computation. Convolutional neural network has the ability of representational learning and can categorize input information according to its hierarchical structure with translation invariant, so it is also known as "translation invariance artificial neural network".

The core ideas of convolution include local receptive field, weight sharing and time or space subsampling to obtain a certain degree of displacement, scale and deformation invariance. The main structures of the convolutional neural network are input layer, convolutional layer, fully connected layer, pooling layer, output layer, etc.

Based on the study of cells in the cat's visual cortex by Huber [7] et al., the convolutional neural network is a specially designed artificial neural network with multiple hidden layers, which is constructed by imitating the biological brain skin layer. Convolutional layer, down-sampling layer and activation function are important components of a convolutional neural network. In this section, we will introduce the theory of convolution neural network and related experiment.

According to the input, weight, excitation function and bias value, the output of the artificial neural node is calculated. A large number of artificial neural nodes are connected with each other to form a neural network. The input value is processed by each layer and the final result is obtained. In this process, the excitation function nonlinearizes the model, which helps to extract the abstract features of the matrix. Using the results and answers, an objective function describing the error is established to calculate the gradient. By the gradient and back propagation formula, the weight and bias of the network are optimized, and the function of the system is mined. Artificial neural network extracts the features from the original matrix by a simple nonlinear model, and only a small amount of manpower is needed in the whole process.

As we all known, convolution neural network adds convolution operation, down-sampling operation and softmax operation on the basis of artificial neural network. In order to obtain different feature maps, convolution filtering is made with multiple convolution kernels. To prevent overfitting, the maximum or average value of each part of the picture represents the characteristics of that part. Meanwhile, after several convolution and pooling operations, the data is input into the softmax layer and the vector output is obtained to achieve multiple classification.

Convolutional neural network has two characteristics: local perception and parameter sharing [8]. Local perception means that each neuron does not need to perceive all the value in the martrix, only perceives the local value of the martrix, and then merges local information at a higher level to get all the feature information of the martrix. The nerve units of different layers are connected locally, that is, the nerve units of each layer are only connected with the part of the former layer. Such a local connection mode ensures that the spatial local mode of the learned convolution check input has strong response. The structure of weight sharing network makes it more similar to biological neural network. Weight sharing reduces the complexity of network model, thus reduces the computational complexity and the number of parameters. This network structure is highly invariant to translation, scaling, skew or other forms of deformation.

Combined with the theory above, we use the improved network architecture based on LeNet-5 to mine system functions. First, the expression level of 100 characteristic genes extracted from the hierarchical clustering process is converted into a 10×10 matrix, and the corresponding answer vector is created. Then the convolutional neural network is used for training and identification. In this experiment, the neural network structure is input layer to convolutional layer to down-sampling layer to convolutional layer to down-sampling layer to full connection layer to output layer. The sizes of convolution kernels are $3 \times 3 \times 6$ and $3 \times 3 \times 12$, and the average operator of 2×2 is used in the two down-sampling layers. At one time, 8 sets of data are selected for training.

The experimental results showed that the time required for each 100 training sessions was about 20 seconds. After more than 200 times of training, the average correct rate of the 20 tests can reach over 85%. However, increasing the number of training times cannot significantly improve the correct rate, and the correct rate always fluctuates around 85%.

IV. Simulation

In this paper, the biological data of primary bladder cancer was selected as an example. The hierarchical clustering method was used to extract systematic dynamic

structure information, while the convolutional neural network technology was used to carry out systematic function mining. The research methods and fields are relatively new, which have the following advantages.

The feature genes of differentially expressed genes were extracted by hierarchical clustering algorithm, and the dimensionality reduction with less information loss was realized. As a result, the subsequent computation was reduced and the computational difficulty was reduced. Meanwhile, by using the convolutional neural network to train the characteristic matrix, the training process and the judgment process are fast, and the judgment accuracy is also high. This paper combines the advantages of the two algorithms, optimizes the structure of the system, and realizes the combination from classification to function judgment, which has strong theoretical significance and wide application prospect.

However, at the same time, the training of the convolutional neural network may fall into the local optimum, which may require multiple training to get out of this state, resulting in a certain degree of decrease in the training efficiency. Furthermore, the clustering of hierarchical clustering is only based on the distance between the samples. For some data, the distance may have special meaning after reaching a certain value. At the same time, some special values with large distances from other samples may affect the structure of the cluster tree.

Therefore, we can make the following improvements. First of all, the specific application of hierarchical clustering should be combined with the background knowledge of the discipline and give some special samples or distance special processing. Secondly, after standardization, the classification basis of different clustering is combined as the basis of improved clustering. For example, the sum of Euclidean distance and trigonometric cosines with a certain weight is used as the clustering basis. Thirdly, the transpose convolution operator is introduced, that is, the rows and columns of the original matrix are added to be 0 before the convolution operation, so that the size of the matrix remains the same or becomes larger after the convolution. This algorithm can make the matrix (image resolution) meet the requirements of the algorithm in the operation of segmentation. Finally, change the fixed step size to a variable step size. The value of the residual error is used as the step size control factor. The larger the residual error is, the longer the step size will be, which will accelerate the training to some extent. Meanwhile, larger stride length and smaller step length can be used in the early training cycle, which is helpful to prevent the parameters from falling into local optimum.

In view of the two technologies mentioned in this paper, hierarchical clustering and convolutional neural network, which have excellent properties and be widely used, are playing an important role in the era of big data. In particular, convolutional neural network technology has obvious advantages in image processing. We also introduced the convolutional neural network to process MNIST data set and conduct image recognition experiments in the following paper we expanded.

There is no doubt that the application of hierarchical clustering and convolutional neural network is not limited to biological data, but can also be applied to computer technology, finance, medicine, natural language and other fields. For example, Vaillant [9] et al. proposed the application of convolutional neural network to face detection firstly. On this basis, Ouyang [10] et al. proposed a joint convolutional neural network-based pedestrian detection model. The model integrates feature extraction, deformation processing, occlusion processing and feature classification into a single convolutional neural network, automatically establishes the relationship between pedestrian parts through end-to-end training, and enhances the separability of features. Li [11] et al. proposed a pedestrian detection method based on memory neural network. First, scan the image to be detected and convert the single image into an image sequence. Second, the convolutional neural network is used to extract features from each image in the sequence to obtain sequence features. Then, the sequence features output a mask map of pedestrian location through a complex convolutional neural network that learns and remembers pedestrian sequence patterns. Finally, the appropriate window is used to frame the pedestrian on the mask diagram to complete the pedestrian detection. This shows that the application of hierarchical clustering and convolutional neural network technology is very extensive, which is worthy of our further study.

V. Conclusion

In this paper, primary bladder cancer was taken as a typical example, the structure of the system was extracted by hierarchical clustering method, while the function of the system was mined by convolutional neural network technology. Using the method of clustering, the stratification boundary of biological data was determined at first and the study data was layered according to the characteristics of biological data. Then, we carried out stratification analysis at different levels, the correlation intensity of cell function was analyzed, the effective discrimination and control of confounding bias was realized, and the structural characteristics of biological data was extracted. Finally, we built complex system structure analysis model and designed relevant algorithm. Combined with the actual data, we carried out the feasibility and comparative study of related theories and methods to realize the perfect combination of theory and practice. The application and expand of technology were also mentioned.

Acknowledgment

This work was supported by the National Natural Science Foundation of China (Grand No. 11371174) and Undergraduate Innovation and Entrepreneurship Training Program (Grand No. 201910295072).

References

[1] Eisen M B, Spellman P T, Brown P O, et al. Cluster analysis and display of genome-wide expression patterns. Genetics, 1998, 95(25): 14863-14868.

[2] Guha S, Rastogi R, Shim K. CURE: an efficient clustering algorithm for large databases. Information Systems, 2001, 26(1): 35-58.

[3] X. Tang and P. Zhu, "Hierarchical Clustering Problems and Analysis of Fuzzy Proximity Relation on Granular Space," in IEEE

Transactions on Fuzzy Systems, vol. 21, no. 5, pp. 814-824, Oct. 2013, doi: 10.1109/TFUZZ.2012.2230176.

[4] Müller A T, Hiss J A, Schneider G. Recurrent Neural Network Model for Constructive Peptide Design. Journal of Chemical Information and Modeling, 2018, 58 (2): 472 -479.R. Nicole, "Title of paper with only first word capitalized," J. Name Stand. Abbrev., in press.

[5] Tran N H, Zhang X, Xin L, et al. De Novo Peptide Sequencing by Deep Learning. Proceedings of the National Academy of Sciences, 2017, 114 (31): 8247 -8252.

[6] Yang B, Liu F, Ren C, et al. BiRen: predicting enhancers with a deep -learning based model using the DNA sequence alone. Bioinformatics, 2017, 33 (13): 1930 - 1936.

[7] Olshausen B A. Emergence of simple-cell receptive field properties by learning a sparse code for natural images[J]. Nature, 1996, 381(6583): 607-609.

[8] Won Y, Gader P D, Coffield P C. Morphological shared-weight networks with applications to automatic target recognition[J]. Neural Networks, IEEE Transactions on, 1997, 8(5): 1195-1203.

[9] Vaillant R, Monrocq C, LeCun Y. Original approach for the localisation of objects in images [J]. Vision, Image and Signal Processing, 1994, 141(4):245-250.

[10] Ouyang Wanli, Wang Xiaogang. Joint deep learning for pedestrian detection [c]//Proc of IEEE International Conference on Computer Vision. 2013:2056-2063.

[11] Li Xudong, Ye Mao, Liu Dan, et al. Memory-based object detection in surveillance scenes [c]//Proc of IEEE International Conference on Multimedia and Expo. 2016:1-6.