MEDS Public Health and Preventive Medicine (2022)
Clausius Scientific Press, Canada

DOI: 10.23977/phpm.2022.020206
ISSN 2616-1915 Vol. 2 Num. 2

Selection of tumor characteristic genes based on data mining technology

Mingxi Chen[1], Yunhao Liu[2], Junming Hou[3, *]

[1]*Shaanxi University of Chinese Medicine, Xixian Avenue, Xixian new area 712046, Shaanxi Province*

[2]*Affiliated Hospital of Shaanxi University of Chinese Medicine, No. 2, Weiyang West Road, Qindu District, Xianyang City 712000, Shaanxi Province, Department of Surgical Thoracic*

[3]*Affiliated Hospital of Shaanxi University of Chinese Medicine, No. 2, Weiyang West Road, Qindu District, Xianyang, Department of Surgical Oncology*

*Corresponding author

Keywords: Data mining technology, Tumor characteristics, Gene selection

Abstract: Gene chip technology is widely used to study gene expression patterns of cells at genome level because it can quickly measure the expression levels of thousands of genes at the same time. Gene microarray technology can track and monitor tens of thousands of gene expression levels in different tissues. It not only provides a powerful scientific basis for cancer biology research, but also helps the classification and identification of cancer tissues. With the wide application of microarray technology in the research of tumor diseases, a large number of tumor gene expression profile data with high dimensions and few samples have been produced. Because of its high efficiency and high throughput, DNA microarray technology has been widely used in various biomedical researches, which can detect a large number of tumor gene expression. Based on data mining technology, today, big data technology has been widely used in all walks of life, which has greatly promoted the development and progress of society. Therefore, in-depth research and discussion on big data technology is of great significance for its future optimization and development.

1. Introduction

In 1985, American scientists took the lead in putting forward the Human Genome Project, aiming at elucidating the sequence of 3 billion base pairs in the human genome, discovering all human genes and figuring out their positions on chromosomes, deciphering all human genetic information, and enabling human beings to fully understand themselves at the molecular level for the first time [1]. With the continuous progress of today's society and the vigorous development of science and technology, the level of life and study of human beings is gradually rising, and at the same time, the degree of attention to health and the awareness of defense against diseases are gradually enhanced [2]. With the implementation and successful completion of the Human Genome Project, a large amount of biological data has been produced, which needs to be analyzed and explained by people in various ways from different angles to gain a deeper understanding and knowledge of life phenomena

[3]. More and more researchers use gene chip technology to obtain the expression levels of thousands of genes from tissue samples [4]. Therefore, gene chip technology has been widely promoted. Compared with the massive accumulation of biological data, our information and knowledge are growing slowly [5]. Medicine, medicine, agriculture and environmental protection urgently need to get useful knowledge from these raw data, which leads to the emergence of bioinformatics, an interdisciplinary subject [6]. In recent years, due to the increasing and accumulation of massive gene expression profile data, the research and analysis of the analysis methods, theories and technologies of these massive data, as well as the characteristic gene selection method and characteristic gene selection method, how to select the best characteristic gene by selecting the appropriate characteristic gene selection method, and then establish an accurate and effective classification model, have become the vital contents in the inquiry category of bioinformatics [7].

2. Feature gene selection method

2.1 Winding method

Entanglement method embeds the construction of classification algorithm into the search process of feature subset, and generally regards classification accuracy as the evaluation criterion of feature importance [8]. In order to search all feature spaces, related algorithms are "entangled" in the classification model [9]. Wrapper method (Wrapper method): Wrapper method was put forward as early as 1990s. This algorithm often combines feature selection process with supervised classification, and the initial target feature set is empty [10]. According to the given feature evaluation criteria, the best one is selected from the original feature set and added to the target set, and the best one is selected in each subsequent iteration process until the target set is optimal and satisfied [11]. The single feature subset space increases exponentially with the increase of the number of features, and some heuristic algorithms are often used for optimization, such as exhaustive iterative search and random search [12]. The advantage of the entanglement method is that it takes into account the interaction between feature subset search and classification algorithm and the relationship between features, but its disadvantage is that the amount of calculation is too large, especially when constructing classification model, the calculation cost is high. Since the final selected feature subset is still to be used for the subsequent learning algorithm, the performance of the learning algorithm we selected is obviously the best evaluation standard, so the performance of the learning algorithm is used as the evaluation standard for the feature selection of the winding method. Sierra et al. compared and analyzed the classical entanglement search algorithms such as sequential forward search, sequential backward search and floating search by using three gene expression profile data sets.

Compared with the filter algorithm, the Wrapper algorithm has higher computational complexity and is difficult to implement. It is unrealistic to use a complex classifier for feature selection, but theoretically, the Wrapper algorithm achieves better results. That is because the Wrapper algorithm combines the feature selection process with the classification process [13]. Although the wrapping method has higher computational complexity, there are still many excellent feature selection methods based on the wrapping method. Sharma et al. proposed a continuous feature selection algorithm. The features are divided into several independent subsets by the successful feature selection (SFS) method. According to the classification performance of subsets, the best features in each subset are selected to get the best feature subset. The selection methods of entanglement information genes mainly include: the combination of neural network and random features; GA/KNN algorithm combining genetic algorithm with KNN algorithm; Step-by-step optimization algorithm is also an information gene selection method that depends on the wrapper method. Wanderley et al. have constructed a new gene screening method combining genetic algorithm and winding method. This paper expounds that nonparametric method is a good choice for sparse data, such as biological gene data. It doesn't need to make any assumptions, and all information comes from the data itself.

2.2 Embedding method

The embedding method fuses feature selection in the training process of the model, and calculates the weight coefficient of each feature. For example, in the process of building branches of decision trees, the embedded feature selection method is adopted, and the core idea is to sort features according to a certain measurement index. Mbedded method (embedding method) combines classifier design with gene screening. At present, gene selection methods based on embedding method mainly include support vector machine and decision tree. The advantage of this method is that it combines the influence of classification algorithm on results, and it is better than filtering method in time and space consumption. Capobianco et al. pointed out that reducing the dimension of gene data is an effective preprocessing method, and the gene processing system can be simplified into several independent parts, and proposed a gene selection algorithm combining embedding method and entropy theory, which can effectively reduce the influence of noise. The feature selection process of vector machine (SVM) is to calculate the weight of each feature, and delete those features with small weight coefficients in the design process of SVM, which is one of the typical classifiers. Wang et al. proposed a first-order generalized learning classification algorithm, which is divided into positive and negative class selection criteria. The information gain is used to select the attribute with the largest contribution rate to generate the criterion. With the addition of data, the training set covered by the criterion will be deleted, and the attribute value with the largest contribution rate will continue to be selected from other training sets for iterative operation. SVM-RFE method is to design a classifier by recursive method to delete genes with small weights. This algorithm has been applied to leukemia data sets and colon cancer data sets, and two and four information genes have been selected, respectively. The classification effect is relatively good. Reich et al. put forward a perturbation iterative gene selection algorithm based on the idea of embedding method, which adopts the method of reverse elimination of redundant features. The algorithm relies on the influence of noise on the classification performance of each feature, that is, if the classification performance changes greatly after adding noise, the feature is a relevant feature.

The embedding method itself is unstable, and in order to avoid this phenomenon, new methods are constantly being produced. Tang et al. use the two-step SVM-RFE method to screen genes. The first step is to use SVM-RFE method to remove irrelevant genes, noise genes and redundant genes according to different screening factors, so as to obtain different subsets; Secondly, the obtained subsets are combined, and the SVM-RFE method is used for gene screening again. In order to overcome the imbalance of some gene microarray data (the number of samples in some categories is much larger than that in others), Anaissi et al. constructed an improved random forest embedding method. Combining with the idea of information fusion, this algorithm first calculated the maximum error cost of each category, and then selected the relevant features by using the random forest method, which showed good performance on Leukemia gene data sets. Weston et al. applied gradient descent method to the design of SVM, which made the obtained characteristic genes have better discrimination ability.

3. Selection of tumor characteristic genes based on data mining technology

3.1 Data mining

With the development of network technology, the use of data mining technology in Web prefetching has gradually become a mainstream trend. Data mining technology can make users search for access models conveniently, and can also establish information related to users by means of cluster analysis, so that users' behaviors on each page can be mastered, and thus more efficient information services can be brought to users. People can't live without the Internet, which plays an important role in promoting the whole world. More and more users have entered the Internet world, so the online world is becoming more and more crowded. The advantage of big data technology is that it can optimize the processing and management of a large amount of data, but it is not without defects. Its defects are that the accuracy of data search is insufficient in the actual operation process, and it is difficult for users to use data. The traditional way can't guarantee the speed of the browser and improve the user's experience. Therefore, according to the current market research, we can classify the user's requests and clarify the user's needs in the process of surfing the Internet. However, at present, how to use data mining technology perfectly in Web prefetching and fully show the value of data mining technology has become the main task of related workers. At present, the optimization and development direction of big data technology lies in improving the accuracy of data search, reducing the difficulty of user data use, and optimizing the data editing process. Users can have a good experience if they use Web pages when they live and work on the Internet. In this way, users can download these pages by caching during the process of accessing them, which reduces the waiting time of user requests.

When modeling, it is necessary to use data mining technology association algorithm to make accurate calculation for the preprocessed data, and then mine the pattern set in the data. Nowadays, big data technology has been widely used in all walks of life, which has greatly promoted the development and progress of society. Therefore, in-depth research and discussion on big data technology is of great significance for its future optimization and development. On the basis of preliminary research on Web technology, this technology can play an important role in the development of data mining technology and improve users' good experience. Data mining technology is the product of the development of computer technology. Using this technology, we can dig out hidden data from large-scale data, and provide important reference value for scientific and technological decision-making. Then, through the obtained results, we can master the behavior of users.

3.2 Gene selection of tumor classification characteristics based on gene expression profile data

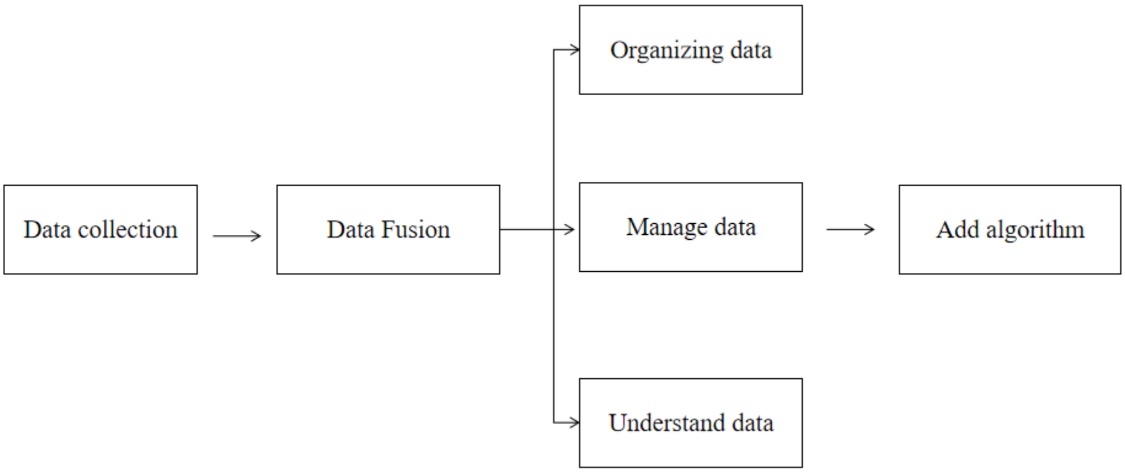

Figure 1: Flow of gene expression data analysis

Gene expression data has the characteristics of small sample, high dimension and high noise, and there are only a few genes related to diseases. Therefore, the existence of a large number of redundant genes makes the expression difference between diseased samples and normal samples small, and makes little contribution to classification. The emergence of gene expression profiles poses new challenges to the two core tasks of bioinformatics: (1) organizing and managing data, and (2) analyzing and understanding data. Instead, redundant genes will greatly increase the running time of the algorithm, reduce the mining performance and greatly increase the search space. Genes are the material basis of inheritance. Every cell in life contains a complete genome, but only some genes can be expressed. The analysis flow of gene expression data is shown in Figure 1.

Because gene expression profile is the measurement of the whole genome expression of a sample by using DNA chip, it is necessary to establish an effective storage and management mechanism for this kind of data, so as to facilitate its quick retrieval. Through gene microarray technology, a large number of biomolecules can be detected and analyzed at the same time, so that massive data can be obtained. Under certain experimental conditions, these massive data can be converted into gene expression data. The gene expression data has high dimensions, small samples and high noise, and only a few genes are related to diseases. Therefore, there are a large number of redundant genes in the gene expression data, which make no contribution to gene classification, but will increase the time and space complexity of search, and even affect the performance of the algorithm, making the results unreliable. A more important problem is how to analyze these data effectively and discover the hidden information and knowledge. This is also the core of bioinformatics research, and has become a research hotspot in the field of bioinformatics. It is urgent to apply gene chip technology to diagnose and type cancer, and to analyze gene expression data. Although gene expression profile data has experienced long-term development, there are still many problems to be solved. Biological sample data has its own complexity, and gene expression profile reflects the expression information of almost the whole genome in individual tissues and cells. It is a "panoramic" record of gene expression information of cells, reflecting a large number of information of different levels of cells. The noise and outliers that may appear in the experiment, as well as errors and identification errors in the data processing process, require researchers to select and study robust "denoising" methods. The framework of gene selection is shown in Figure 2.

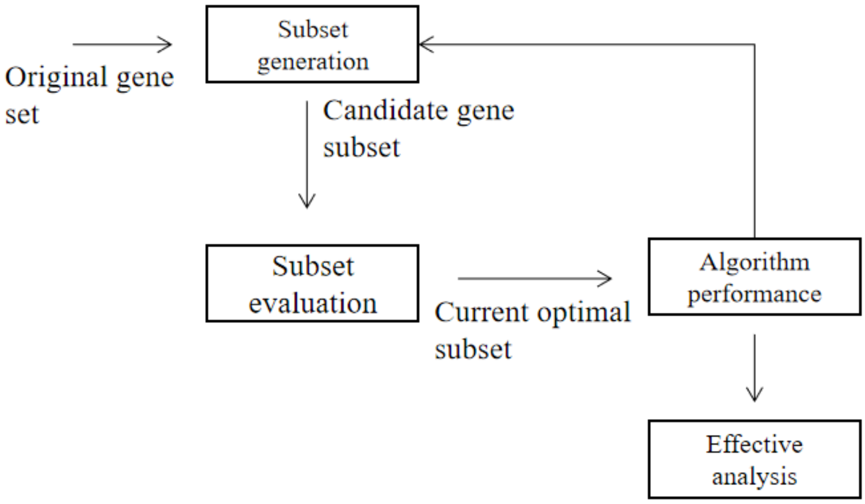

Figure 2: Characteristic gene selection framework

How to effectively analyze and mine the information contained in it according to the goal and draw new biological rules is also a challenge to the traditional machine learning methods. A notable feature of gene expression profile data set is that there are few samples, high dimensions, and a large number of complex relationships among sample attributes. A large number of experimental studies show that the gene expression profile data increases exponentially, and the number of genes is much larger than the number of experimental samples, so the problem of dimension disaster is also a big problem faced by biology. From the biological point of view, only a few genes are really related to the phenotype of the sample. Therefore, how to find out the genes that affect the phenotypic information of the sample, that is, the information genes or the classified characteristic genes of the sample, becomes the key and the main difficulty of tumor gene expression profile analysis.

4. Conclusions

Nowadays, cancer has become the number one enemy of human health risks, so it is urgent to study the early diagnosis and drug development of cancer. With the continuous development of gene chip technology, gene expression profiling technology has been widely used in genome analysis, and it is also a major revolution in the research of gene function. Gene microarray technology can analyze tens of thousands of genes at the same time, changing the traditional experimental technology that only one gene expression can be studied at a time. With the development of microarray technology, gene expression profiling technology has been widely used in genome analysis. Microarray technology can analyze thousands of genes at the same time in one experiment, changing the traditional experiment that only one or several genes can be analyzed at a time. The massive data produced by microarray technology at the same time has the characteristics of high dimension, small sample size and many redundant genes. How to select a small number of characteristic genes with strong recognition ability from these huge data poses a severe challenge to some traditional machine learning methods.

References

[1] Imtiyaz H Z, Williams E P, Hickey M M, et al. Hypoxia-inducible factor 2α regulates macrophage function in mouse models of acute and tumor inflammation [J]. Journal of Clinical Investigation, 2018, 120(8): 2699-2714.

[2] Francesco S, Fabio P, Mariann M, et al. TrAp: a tree approach for fingerprinting subclonal tumor composition[J]. Nucleic Acids Research, 2019, 41(17): e165-e165.

[3] Jackson T L. A mathematical model of prostate tumor growth and androgen-independent relapse [J]. Discrete and Continuous Dynamical Systems - Series B (DCDS-B), 2017, 4(1): 187-201.

[4] Bonnard I, Rolland M, Salmon J M, et al. Total Structure and Inhibition of Tumor Cell Proliferation of Laxaphycins [J]. Journal of Medicinal Chemistry, 2017, 50(6): 1266-1279.

[5] Maria Ibáez-Vea, Zuazo M, Gato M, et al. Myeloid-Derived Suppressor Cells in the Tumor Microenvironment: Current Knowledge and Future Perspectives [J]. Archivum Immunologiae et Therapiae Experimentalis, 2017, 66(2): 1-11.

[6] Huang J K, Jia T, Carlin D E, et al. pyNBS: a Python implementation for network-based stratification of tumor mutations [J]. Bioinformatics, 2018(16): 16.

[7] Epardaud M, Elpek K G, Rubinstein M P, et al. Interleukin-15/interleukin-15R alpha complexes promote destruction of established tumors by reviving tumor-resident CD8+ T cells. [J]. Cancer Research, 2017, 68(8): 2972-2983.

[8] O'Toole C, Stejskal V, Perlmann P, et al. Lymphoid cells mediating tumor-specific cytotoxicity to carcinoma of the urinary bladder. Separation of the effector population using a surface marker. [J]. Journal of Experimental Medicine, 2019, 139(3): 457-466.

[9] Mcgranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future[J]. Cell, 2017, 168(4):613-628.

[10] O'Hara S Mark, Moreno J G, Zweitzig D R, et al. Multigene Reverse Transcription-PCR Profiling of Circulating Tumor Cells in Hormone-Refractory Prostate Cancer [J]. Clinical Chemistry, 2020(5): 5.

[11] Galbraith G M, Steed R B, Sanders J J, et al. Tumor necrosis factor alpha production by oral leukocytes: influence of tumor necrosis factor genotype. [J]. Journal of Periodontology, 2017, 69(4): 428-33.

[12] Davoli T, Uno H, Wooten E C, et al. Tumor aneuploidy correlates with markers of immune evasion and with reduced response to immunotherapy [J]. Science, 2017, 355(6322): eaaf8399.

[13] Roit F D, Engelberts P J, Taylor R P, et al. Ibrutinib interferes with the cell-mediated anti-tumor activities of therapeutic CD20 antibodies: implications for combination therapy [J]. Haematologica, 2017, 100(1): 77-86.