Application Analysis of Feature Selection (FS) in Bioinformatics

Chang-xin Song
Qinghai Normal University, Qinghai, China

Abstract. The feature selection (FS) algorithm selects the effective data from the M original features. The effective feature of low-dimensional data is the structural optimization of the high-dimensional samples. The sample data of bioinformatics was studied using the feature selection (FS) algorithm, which was mainly applied to gene sequence analysis, oligonucleotide array analysis and proteome data analysis. This will explore the intrinsic characteristics of biological systems. This paper mainly discusses the application of feature selection (FS) algorithm in life informatics, and points out the development direction of FS algorithm in bioinformatics by analyzing the problems existing in FS algorithm.

1. Introduction
Bioinformatics has a large number of Multi-latitude sample data. Statistical model or syntactic identification cannot be adapted to the characteristics of large volume sample data. Therefore, introduce the feature selection algorithm is particularly important. In the feature selection (FS) algorithm, the validity of all the obtained feature sets can be validated in order to improve the accuracy of the results. Therefore, the establishment of the bioinformatics feature selection (FS) model can effectively promote the development of medical science such as biological information.

2. Feature selection (FS) overview and model algorithm

2.1. Definition and overview of feature selection (FS)
Feature selection is the optimization of specific indexes in the system to reduce data changes in data information search. FS algorithm is the key data preprocessing step in pattern recognition. It can analyze and improve the information characteristics of bioinformatics, which is able to satisfy the needs of statistics and analysis of multi-latitude data. Feature selection needs to select the original characteristics that can describe the biological mode from all data information and reduce the data dimension of its characteristics. In the case of high-dimensional and sample size, the applicability of the original statistical theory is reduced, and the lower dimension is beneficial to the exploration of its internal structure. The feature selection FS is the process to extract raw information from the whole model sample, then to classify and filter it. The main purpose of this process is data optimization. Feature selection is one kind of the feature dimension reduction, which is used to reduce the dimension and remove the method of fitting phenomenon. Feature reduction divides into feature selection and feature extraction. The feature selection selects the effective feature subset, which means remove uncorrelated or redundant features. The purpose of feature selection is to reduce the number of features, improve the model precision and reduce the running time. Feature selection, which can clearly show the final descending dimension result, has no change in the characteristics value after selection.
2.2. The evaluation criterion classifies feature selection (FS) method

The establishment of feature selection (FS) is to detect the adaptive capability of the machine sample, and to excavate and analyze the potential data information. Feature selection (FS) can actively look for data content that matches biological information. In the process of the original feature reduction, the accuracy of the data model is guaranteed. Essence of the dimensionality reduction is the study of mapping function $f: x \rightarrow y$. $X$ is the original data expression, $y$ is the low dimensional vector expression after the data point mapping. Usually the dimension of $y$ is less than the dimension of $x$. Feature selection (FS) is a subset search for the original feature, which will establish the feature model through the introduction of additional complexity.

2.2.1. Wrapper method

Wrappers use a search algorithm to search for all possible subsets of feature spaces and evaluate the operation of each feature space subset. After determining the feature subset search procedure, the determination and evaluation of the required subset can be completed. The wrapper method needs to establish a large classifier, which is beneficial to the accurate classification selection of the characteristic subset. However, the Wrappers are computationally time-consuming, and there is a risk of fitting.

2.2.2. The filter method

The filter method is similar to Wrappers, it uses a simpler filter to evaluate the feature space subset rather than using the model. Variable sorting is a typical filtering method, but it ignores the possible interdependencies among the features. The feature subset of the filter method contains all the characteristics, and the possible connection between the characteristics is the primary way to determine the data characteristics, and the classifier selection cannot affect the final evaluation result.

2.2.3. The embedding method

The embedded method combines feature selection in the process of model training. For example, the decision tree uses the embedding method in the process of branching, and the characteristics are also sorted according to a metric. The embedding method is a linear combination of the original features by classifier, and the combined characteristics are used as the new feature input into the training model. It is the best feature subset. Compare with other post-merger data search algorithms, the variables in this algorithm are widely associated. There are significant differences in the evaluation criteria of different search stages. The embedding method can be selected directly from the original feature, and the characteristics can be transformed and combined to form new available features.

| Model                  | Advantages                                      | Disadvantages                                      |
|------------------------|-------------------------------------------------|---------------------------------------------------|
| Univariate filter      | Run fast; quick update iteration; Can be run separately from the classification model | Ignore the linkage between feature sets; Ignore the interaction between the classifiers |
| Multivariate filter    | High dependency between models; Strong computing power; Can be run separately from the classification model | Slow operation; Ignore the interaction between the classifiers |
| Wrapper certainty      | High dependency between models; Low computational density; Interact with the classifier | Overfitting risk; The classifier has highly interdependent |
| Wrapper randomness     | High dependency between models; Adopt global optimization scheme; Interact with the classifier | Overfitting risk; Large computation; Classifiers are highly dependent on features |
| Embedding method       | High dependency between models; The computing power is better than the wrapper method | Large computation; The classifier is determined by the selection characteristics |
3. The application of feature selection (FS) in bioinformatics

3.1. Microarray Analysis

Microarrays are also known as oligonucleotides arrays. It integrates the known sequences on a solid surface. The array of nucleic acid sequences that are tagged in a biological cell or tissue is interbred with the probe array. After detecting the interbreeding probe of the corresponding position, the rapid detection of genetic information is completed. The microarray has small single data and large sample size, it has a great advantage in bioinformatics research. The oligonucleotide array mainly analyzes the high dimensional data. It uses single variable, multivariable filter, wrapper, etc., to expand the feature selection (FS) analysis of bioinformatics. The single variable technology is widely used in the field of oligonucleotide array, which is the dominant technology of microarray data collection and evaluation.

The main sample test technology which used in microarray is T test and variance analysis. T test has a good test and analysis capability for the small sample data sets and the noise in the collection. The characteristics of the potential gene distribution were analyzed through the sum of squares and ratios, rank and test of the group and the nominal value of the torque feedback torque. The above three feature selection tests have low dependence on the data hypothesis and can arrange the small sample data in the oligonucleotide array. The resistance of external interference is also very strong. Multivariate technical analysis includes Markov Model, completely fair scheduler algorithm, feature selection MRMR algorithm and so on. Wrapper will search all feature space subset, and through the receiver-operating characteristic curve to represent the final result. Moreover, the ROC curve can also evaluate all the test result analysis.

3.2. Mass spectrometry (ms)

The separation and labeling of ions is the main way to diagnose diseases and is also the major method of protein biosynthesis markers. In mass spectrometry, a large number of data sets are required to be processed, which contains more useless noise. Therefore, the application of feature selection algorithm can effectively solve the influence of other factors on data collection and it will improve the computational accuracy of the algorithm. Single - variable encapsulation is the most widely used feature selection (FS) algorithm in mass spectrum analysis, it is used for random searching through genetic algorithm and particle swarm optimization algorithm. The genetic algorithm searches the optimal solution by simulating the natural evolution process. Especially, we need to start from the peak value or character subset of the data set. Mass spectrometry analysis often uses multiple classifiers for embedding, which can solve the high-dimensional problem in large feature space by using the weighted support vector machine. The classification of neural network classifier has high accuracy and high parallel distribution processing ability. The decision tree algorithm is simple and easy to understand, which is suitable for handling samples with missing attributes, and can carry out a large number of data source analysis in dealing with unrelated features. The current cancer diagnostic scheme is multivariate graphical feature fusion scheme which can perform the dimensionless work of high-dimensional model mapping. The protein-mass spectrometry in the high dimensional data area is entered into the multivariate map region through the mapping region, thus completing the information marking and reduction processing. Multi-level hierarchical control structure is used to solve complex decision problems.

3.3. High dimensional sample data analysis

The high dimensional sample data can reduce the accuracy and fitting problem of sample size by evaluating the standard and the accurate selection of feature extraction. Adaboost + and model can complete the processing of small sample data. The Adaboosting algorithm of feature selection can construct sub-classifiers with various methods. The feature subsets collection is beneficial to the classification of data information. AdaBoosting is a highly accurate classifier. The combination of multiple feature selection (FS) algorithm improves the calculation precision and results in more reliable classification results.
3.4. Protein sequence analysis

3.4.1. Content analysis coding
Most of the protein characteristics can be extracted and judged from subseries, and there is a wide association between each subseries. Therefore, it is very important to analyze the implicit parameters in protein sequences by using Hidden Markov Model software. Hidden Markov Model software is used to predict the structure of helical transmembrane proteins by inserting certain characteristics in different orders of characteristics. The feature selection (FS) algorithm mainly analyzes the latent information of data encoding, and analyzes the related characteristics of the sequence by using various classifiers. The second kind of problem of content analysis is to predict protein function from the sequence. The rule repetition prediction in protein polypeptide chain focuses on the function study of the whole molecular level, which can accurately predict the three-dimensional space structure of protein molecules. The algorithm of parallel genetic algorithm based on structure classification and bit matrix coding can solve the complex optimization problem of multi-population parallel evolution.

3.4.2. Single nucleotide DNA sequence polymorphism
Single nucleotide DNA sequence polymorphism is a common phenomenon of human genetic variation. Mutant gene of the chromosome marker position in the single nucleotide can be transmitted by genetic mutation. There has one in every 500 to 1000 base pairs of the variation of single nucleotide polymorphisms, which will has the total of 3 million or more. Therefore, DNA sequence selection of single nucleotide polymorphisms should select a single nucleotide polymorphic subset with sufficient sequence information for the further diseases analysis. The contiguous sites of single nucleotide polymorphisms will combine the chromosomes into a single genetic mutation. The chromosome allele of a single region contains the polymorphic characteristic of single nucleotide DNA sequence that exists in most groups. Single nucleotide DNA sequence polymorphic sites in one region can be represented by single nucleotide polymorphisms, most of the genetic polymorphism patterns can be represented. The current SVM support vector machine, genetic algorithm and other feature selection (FS) algorithm, are the partition marking methods of single structure.

4. Conclusion
The feature selection (FS) algorithm can obtain the biological information in the database through global searching or random searching. The feature selection (FS) can be used to preserve all the original features of the creature, to extract the characteristic subset from the original feature then achieve the health reduction. The effective combination of multiple feature extraction schemes can improve the accuracy and reliability of the feature subsystems.

Reference
[1] Xie Juanying, Qu Ya’nan, Wang Mingzhao. Unsupervised feature selection algorithms based on density peaks, Journal of Nanjing University (Natural Sciences),2016.4.735-745
[2] Xie Hong, Zhou Xiaoli, Xia Bin, Yang Wenlu, Yao Nan. Research on Selection Algorithm of EEG Signal Features and SVM Based Model for Driver’s Fatigue Recognition, Chinese Journal of Biomedical Engineering, 2014.4.p.482-486
[3] Yao Minghai, Wang Na, Qi Miao, Li Yan. Study on feature selection method of modified maximal relevance minimal redundancy. Computer Engineering and Applications, 2014, 50(9):116-122.
[4] Xu Jialin, Zuo Guokun, Motor Imagery Electroencephalogram Feature Selection Algorithm Based on Mutual Information and Principal Component Analysis, Journal of Biomedical Engineering,2016.2
[5] Li Shufang, Research on Red Tide Algae Images Feature Selection Method Based on ReliefF and SBS, Science Mosaic,2014.5,p.12-15
[6] Cao Ying, Comparative study of feature selection algorithm in hierarchical classification, Digital
[7] Zhou Junling, Chen Lin, Xu Baowen, An Unsupervised Feature Selection Approach Based on Mutual Information, Journal of Computer Research and Development, 2012, 2, p. 372-3382