Application of machine learning methods in bioinformatics

Cite as: AIP Conference Proceedings **1967**, 040015 (2018); https://doi.org/10.1063/1.5039089
Published Online: 23 May 2018

Haoyu Yang, Zheng An, Haotian Zhou, and Yawen Hou

Application of Machine Learning Methods in Bioinformatics

Haoyu Yang [a)], Zheng An [b)], Haotian Zhou and Yawen Hou

College of software, Jilin University, Changchun 136000, China

[a)] Corresponding author: yanghaoyujlu@163.com
[b)] anzheng5515@mails.jlu.edu.cn

Abstract. Faced with the development of bioinformatics, high-throughput genomic technology have enabled biology to enter the era of big data. [1] Bioinformatics is an interdisciplinary, including the acquisition, management, analysis, interpretation and application of biological information, etc. It derives from the Human Genome Project. The field of machine learning, which aims to develop computer algorithms that improve with experience, holds promise to enable computers to assist humans in the analysis of large, complex data sets.[2].This paper analyzes and compares various algorithms of machine learning and their applications in bioinformatics.

INTRODUCTION

Machine learning is closely related to computational statistics, which also focuses on prediction-making through the use of computers. It has strong ties to mathematical optimization, which delivers methods, theory and application domains to the field. Machine learning is sometimes conflated with data mining, where the latter subfield focuses more on exploratory data analysis and is known as unsupervised learning. Machine learning can also be unsupervised and be used to learn and establish biology database and help find laws in gene sequences.

PRINCIPAL COMPONENT ANALYSIS

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. Principal component analysis is a statistical method. Through orthogonal transformation, a set of variables that may have correlation is transformed into a set of linearly uncorrelated variables. The transformed set of variables is called principal component. The goal of PCA is to find r, which is used when analyzing complex cubes. When the number of variables is greater than the number of samples, the PCA can minimize the sample dimension to the number of samples without losing the amount of information. [3]It can be seen as a step in the preparation of complex experimental data.The process of PCA is as follows:

- Go to the average, that is, subtract the average for each feature
- Calculate the covariance matrix
- Calculate the eigenvalues and eigenvectors of the covariance matrix
- Sort eigenvalues from largest to smallest
- Keep the largest number of eigenvectors
- Transform the data into a new space constructed by a feature vector

With the help of the machine learning algorithm PCA, we identify five kinds of differentiated alveolar cells and this is shown in Fig.1.

6th International Conference on Computer-Aided Design, Manufacturing, Modeling and Simulation (CDMMS 2018)
AIP Conf. Proc. 1967, 040015-1–040015-4; https://doi.org/10.1063/1.5039089
Published by AIP Publishing. 978-0-7354-1672-7/$30.00

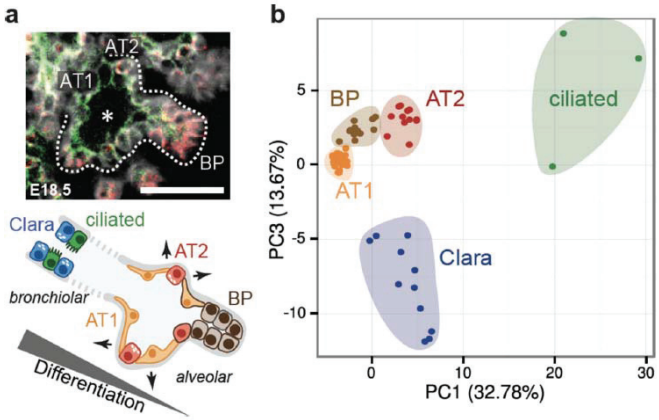

FIGURE 1. Distinguish five kinds of cells by the first and third PCA

BACK ERROR PROPAGATION ALGOTITHM

Artificial neural network is a hot research field in recent years. [4]It has always been one of the important research contents in the field of artificial neural network. The basic principle of BP network model processing information is the input signal X_i acts on the output node through an intermediate node (hidden layer point).After nonlinear transformation, it generates an output signal Y_k. Each sample trained by the network includes an input vector X and a desired output t, the network output Y, a desired output t by the adjustment of the input node,hidden layer node connection strength W_{ij} value, the hidden layer node, the output node connection strength T_{jk} and threshold, so that the error along the gradient direction is the least. After repeated learning and training to determine, the minimum error corresponding to the network parameters (weights and thresholds) and the training stops. At this point, the trained neural network can process the non-linear transformed information with the smallest output error on the input information. The principle of back error propagation is shown in Fig.2.

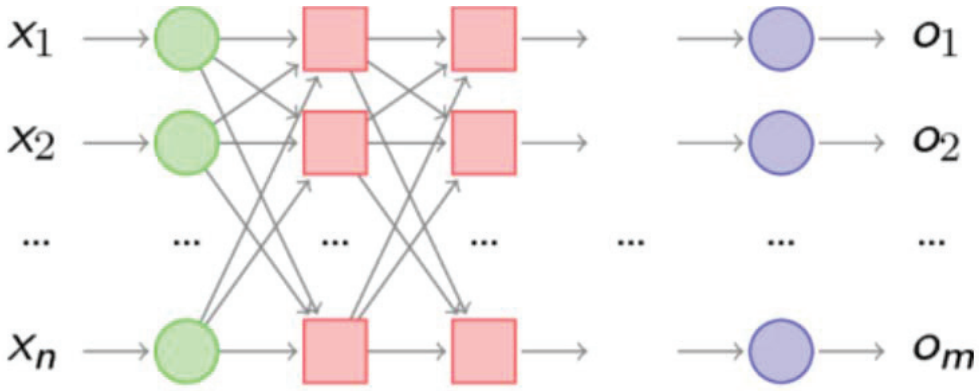

FIGURE 2. The principle of BP

In neural networks, neurons receive input signals from other neurons, which are multiplied by weights and added to the total input values received by the neurons, then compared with the current neurons' thresholds and then processed by an activation function, as a result, it generate neuron output. The ideal activation function is a step function, "0" corresponds to neuron depression, and "1" corresponds to neuron excitation. However, the shortcoming of step function is discontinuous, non-conductive and unsmooth, so the sigmoid function is often used as an activation function instead of a step function.

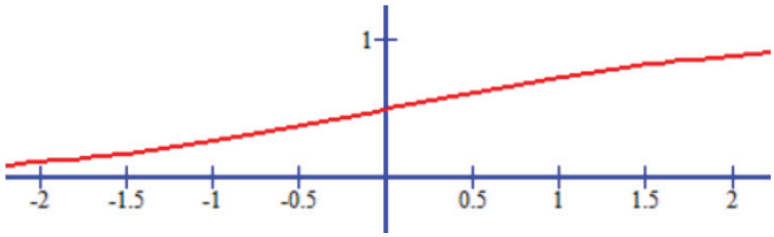

FIGURE 3. Sigmoid function

Protein secondary structure prediction is based on the known primary structure, the prediction methods and techniques to achieve the classification of secondary structure prediction. For BP neural network, the input is the primary sequence of known as the protein, the output is the secondary structure type. In this paper, a 3-layer BP neural network is used as a classifier to predict the protein secondary structure. Design, the BP network input layer is designed to slide along the amino acid sequence of the window, the window position is symmetrical.[5]

The disadvantage of this algorithm is that it can be overfit and needs lots of tricks.

CONVOLUTIONAL NEURAL NETWORK

In machine learning, a convolutional neural network (CNN) is a class of deep, feed-forward artificial neural networks that has successfully been applied to analyzing visual imagery. a deep convolutional neural network (abbreviation as CNN), which has strong enough induction ability.[6] Convolutional networks were inspired by biological processes] in which the connectivity pattern between neurons is inspired by the organization of the animal visual cortex. Individual cortical neurons respond to stimuli only in a restricted region of the visual field known as the receptive field. The receptive fields of different neurons partially overlap such that they cover the entire visual field. A typical convolutional neural network is shown in fig.4.

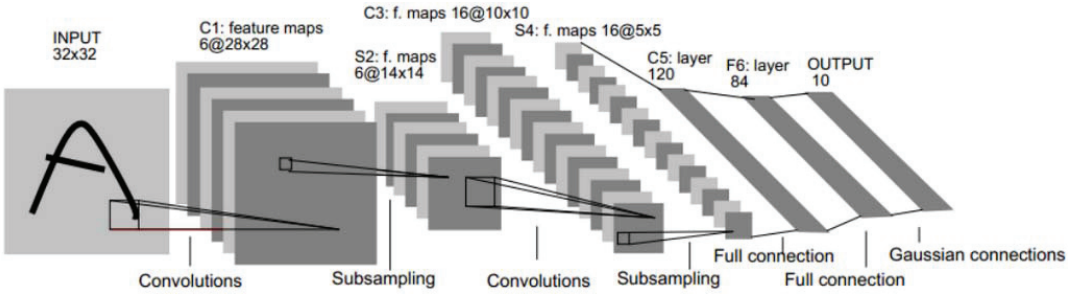

FIGURE 4. A typical convolutional neural network

In bioinformatics, CNN an be used for DNA sequence variation site detection.[6]A deep convolutional neural network can call genetic variation in aligned next-generation sequencing read data by learning statistical relationships (likelihoods) between images of read pileups around putative variant sites and ground-truth genotype calls.

CONCLUSION AND PROSPECT

In conclusion, the machine learning algorithms can be used in bioinformatics .PCA can be help identify different alveolar cells. BP algorithm can be used for predicting the secondary structure of protein. CNN can help DNA sequence variation site detection. Within the field of data analytics, machine learning is a method used to devise complex models and algorithms that lend themselves to prediction.

ACKNOWLEDGMENTS

The research work was supported by the College Students Innovation Training Program Fund for Jilin University under Grant No. 2017B54483.

REFERENCES

1. Ma, Chuang, Zhang, et al. Machine learning for Big Data analytics in plants. [J]. Trends in plant science, 2014, 19(12):798-808.
2. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics [J]. Nature Reviews Genetics, 2015, 16(6):321-32.
3. Zhang Zhen, Li Junli. Machine learning methods and their application in bioinformatics [J] .Journal of Jishou University (Natural Science Edition), 2006, 27 (4): 28-32.D. L. Davids, “Recovery effects in binary aluminum alloys,” Ph.D. thesis, Harvard University, 1998.
4. Huang Li. Improvement and Application of BP Neural Network Algorithm [D]. Chongqing Normal University, 2008.
5. WANG Fei-Lu, SONG Jie, SONG Yang. Application of BP Neural Network in Prediction of Protein Secondary Structure [J] .Computer Technology and Development, 2009, 19 (5): 217-219.
6. Li, Qiaoliang, et al. "A supervised method using convolutional neural networks for retinal vessel delineation." International Congress on Image and Signal Processing IEEE, 2016:418-422.