The Basic algorithm of machine learning and its application in protein phosphorylation

Zheyao Gao[1, a]

[1]School of Electronic Information and Communications, Huazhong University of Science and Technology, Wuhan 430074, China.

[a]3065231608@qq.com

Keywords: machine learning phosphorylation, bayesian decision theory, feature selection, phosphoproteomics.

Abstract. Phosphorylated proteomics is a discipline that studies all phosphorylated proteins involved in the life process and plays an important role in proteomics. With the advancement of technology, there have been many experimental methods for protein phosphorylation and bioinformatics-based computing strategies. The application of data-driven machine learning methods in phosphate-catalyzed research has matured and has become the mainstream method in this field. This paper mainly summarizes the main methods and principles of machine learning, including Bayesian Decision Theory (BDT), Random Forests (RFs), AdaBoost, Support Vector Machines (SVMs)and etc. Finally, the application of these methods in the field of phosphorylation is reviewed in this article.

Introduction

Phosphorylated proteomics is an important topic in the field of bioinformatics today [1]. With the development of biotechnology, there have accumulated a large number of phosphorylation data, but how to dig out the inherent modification law of phosphorylation from these biological data and then explain the mechanism of bioinformatics is an urgent problem that experts need to solve [2]. The machine learning method is an effective tool for solving such problems and this method uses the relevant biometrics to model the training data to classify or return the biological data of unknown tags. There are 518 genes encoding protein kinases in the human genome [3], so it is a typical multi-class classification problem to predict which protein kinase phosphorylation or phosphorylation site corresponds to which kinase. We can transform these multi-class classification problems into two kinds of classification problems. Here, we discuss only two kinds of classification problems. In this paper, the machine learning methods commonly used in the field of phosphorylation research are reviewed, and the corresponding theoretical basis and application in the field of phosphorylation research are introduced.

Machine learning

Machine Learning is to study how to use computer simulation or how to achieve human learning activities. It is another important application field of artificial intelligence after expert system. It is one of the core problems of computer intelligence and the core subject of artificial intelligence research. It is applied in all fields of artificial intelligence [4].

Learning is a process of knowledge acquisition with a specific purpose. Its internal performance is mainly due to the continuous establishment and modification of the new knowledge structure, and the external performance is the improvement of performance. A learning process is essentially a process that the learning system transforms the information provided by the mentor (or expert) into a form that can be understood and applied by the system. According to the degree of system dependency on

the instructor, the learning methods can be categorized as: Rote learning, learning from instruction, learning by analogy, learning from induction, learning by observation and discovery and so on.

In addition, in recent years, learning methods have experienced a visible development based on interpretation, examples, concepts, neural network learning and genetic learning.

Basic algorithm of machine learning

The Theory of Bayesian Decision Theory. An important branch of machine learning is the Bayesian machine learning. The Bayesian method first originated in a special case of the Bayesian theorem by the British mathematician Thomas Bayesian in 1763 [5]. After the concerted efforts of a number of statisticians, Bayesian statistics have been gradually established since the 1950s, making it an important component of statistics. The Bayesian theorem is known for its unique understanding of the degree of subjective confidence in probability. Bayesian statistics have been widely and far-reaching used in the fields of posteriori reasoning, parameter estimation, model detection, hidden variable probability model and so on.

The Bayesian method has many applications in the field of machine learning, which is from univariate classification and regression to multivariate structured output prediction, from supervised learning to unsupervised and semi-supervised learning. The Bayesian method is available to kinds of learning tasks. Such as forecasting tasks:

Given the training data D, the prediction of the future data x is obtained by the Bayesian method:

$$p\left(x\middle|D\right)=\int_{\theta}p\left(x,\Theta\middle|D\right)=p\left(x\middle|\Theta,D\right)p\left(\Theta\middle|D\right) \tag{1}$$

It should be noted that when the model is given, the data is derived from an independent and identical sample, so $p\left(x\middle|\Theta,D\right)$ is usually simplified as $p\left(x\middle|\Theta\right)$

The Theory of Random Forests. Random forest is an integrated statistical machine learning method based on decision tree [6]. There are three kinds of nodes in the decision tree: root node, intermediate node, leaf node. As shown in Figure 1, the directed edge always points from the root node to the intermediate node, and finally points to the leaf node.

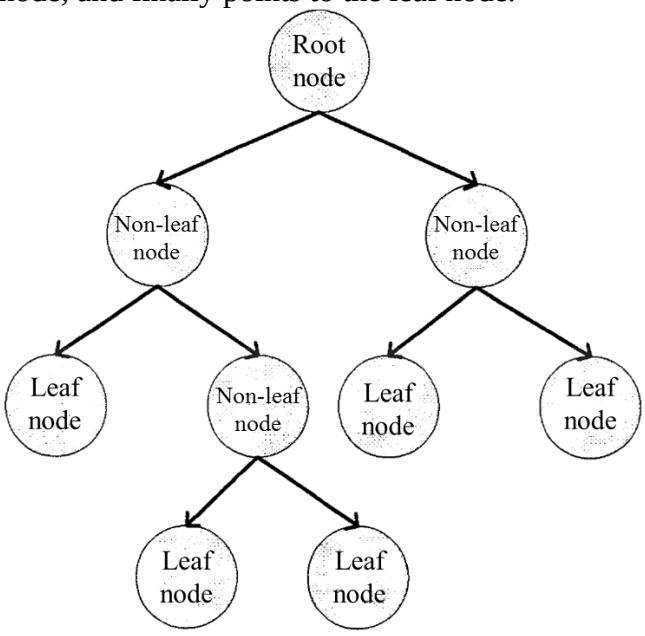

Fig. 1 Decision tree diagram

The Theory of AdaBoost. A number of weak classifiers weightily sum up into the strong classifier, where the weak classifiers meet the correct rate of 50% in the case of two categories, thus, the strong classifier accuracy will increase a lot [7]. The process of AdaBoost is as follows: At the beginning, the same weight is given to the samples of the training dataset, which form the vector D. Then, a weak classifier is trained by these weighted sample data sets and the weighted error rate is calculated. After this, the weight of weak classifier calculated can be calculated by the error rate (the higher the error

rate, the smaller the weight). In the second training, the weight of the sample is adjusted according to the case where the sample is misaligned. The weights are increased by the misclassified samples, and the weights are correctly reduced, so that the trainer will pay more attention to those in the training process is misaligned, and the final selection of the weighted error rate of the smallest classifier is regarded as the round of the weak classifier. Then according to the error rate, the corresponding weight to calculate tweak classifier can be measured. The training is terminated when the error rate is less than the set threshold, and the final classifier is composed of T weak classifiers. The weight of each classifier represents the accuracy of classification.

The Theory of Support Vector Machines. The support vector machine algorithm (SVM) which is come up with Vapnik and other scholars is one of the most influential results in recent 10 years' development of machine learning, pattern recognition and neural network [8]. Support vector machine classification algorithm has four salient features: 1) use the idea of large interval to reduce the VC dimension of the classifier to realize the principle of structural risk minimization and control the generalization ability of classifier; 2) use Mercer kernel to realize the linear algorithm Linearity; 3) sparsity, that is, a small number of samples (support vector) coefficient is not zero, in terms of promotion, the number of support vector in the statistical sense of the corresponding good promotion ability, from the calculation point of view, support vector The computational complexity of the kernel form is reduced; 4) The algorithm is designed as convex quadratic programming problem, which avoids the multi-solution.

Application of machine learning in phosphorylation

The application of BDT in phosphorylation. Xue et al. introduced Bayesian decision theory into kinase-specific phosphorylation site prediction studies [9]. The amino acid frequencies of each pair of positive and negative data sets were counted according to the conservativeness of the sequences. The amino acid frequencies of 9 amino acids were calculated by selecting 8 amino acids in the neighbor phosphorylation sites. They calculated the probability distribution for each position and each amino acid in each sample.

The application of RFs in Phosphorylation. Fan et al. were orthogonal to the local sequence around the phosphorylation site, using the Willows kit to construct a random forest, and the kinase specifically predicted the potential phosphorylation site [10]. The predictive performance at different window widths was evaluated using five-fold cross validation on four typical kinases: CDC2, CK2alpha, MAPK and PKCa datasets.

The application of AdaBoost in Phosphorylation. Cai method was used to select the substrate position and amino acid properties affected by phosphorylation around the phosphorylation site. AdaBoost was then used to select the target and classifier training to developed the kinase-specific phosphorylation site prediction tool AproPhos. Furthermore, AdaBoost is also used to extract the rules, and the understandable amino acid distribution is given, and the prediction results are explained [11].

The application of SVM in Phosphorylation. SVM has a wide range of applications in the field of phosphorylation because of its good classification performance. The research work of the Kim study group was one of the earlier studies to introduce SVM into the prediction of phosphorylation sites. They modeled the data of the four kinase family and the four kinase groups using the sequence information of the phosphorylation site by SVM, and developed the phosphorylation site prediction tool PredPhospho [12].

Conclusion

With the development of proteomics and the progress of high-throughput experimental techniques, the problem of lack of protein kinase information is becoming more and more serious, which also hinders further study of phosphorylation. Using the existing phosphorylation sites and the corresponding protein kinase information, the development of a dedicated machine learning

algorithm can effectively assign the corresponding kinases to the phosphorylation sites obtained for the experiment, thus providing guidance for biomedical research related to protein phosphorylation.

Reference

[1] MR Wilkins, JC Sanchez, AA Gooley, RD Appel, I Humphery-Smith, DF Hochstrasser, KL Williams. Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it [J]. Biotechnology & Genetic Engineering Reviews, 1996, 13 (1): 19-50.

[2] EA Berry, AR Dalby, ZR Yang. Reduced bio basis function neural network for identification of protein phosphorylation sites: comparison with pattern recognition algorithms [J]. Computational Biology and Chemistry, 2004, 28: 75-85.

[3] G Manning, DB Whyte, R Martinez, T Hunter, S Sudarsanam. The protein kinase complement of the human genome [J]. Science, 2002, 298 (5600): 1912-1934.

[4] AL Blum, P Langley. Selection of relevant features and examples in machine learning [J]. Artificial Intelligence, 1997, 97 (1–2): 245-271.

[5] T Bayes. An essay towards solving a problem in the doctrine of chances [J]. London: Philosophical Transactions Royal Society, 1763, 53: 370-418.

[6] L Breiman, J Friedman, R Olshen, C Stone. Classification and regression trees [J]. Chapman & Hall/CRC, 1984, 40 (3): 17–23.

[7] Y Freund, RE Schapire. A desicion-theoretic generalization of on-line learning and an application to boosting [J]. European Conference on Computational Learning Theory, 1995, 904 (1): 23-37.

[8] BE Boser, IM Guyon, VN Vapnik. A training algorithm for optimal margin classifiers [J]. Workshop on Computational Learning Theory, 1992, 5: 144-152.

[9] Y Xue, A Li, L Wang, et al. PPSP： prediction of PK-specific phosphorylation site with Bayesian decision theory [J]. BMC bioinformatics, 2006, 7: 163.

[10] W Fan, L Zou, A Li, et a1. Prediction of protein kinase-specific phosphorylation sites using random forest algorithm [C]. 2012 5th International Conference on Biomedical Engineering and Informatics, 2012, 918-921.

[11] JJ Cai. Study on the prediction and rule extraction of protein phosphorylation site [D]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences, 2006, 1-54.

[12] JY Kim, K Huh, R Jung et al. Identification of BCAR-1 as a new substrate of Syk tyrosine kinase through a determination of amino acid sequence preferences surrounding the substrate tyrosine residue [J]. Immunology Letters, 2011, 135: 151-157.