早稲田大学大学院情報生産システム研究科

博 士 論 文 概 要

論 文 題 目

Study on the Predictions of Gene Function and Protein Structure Using Multi-SVM and Hybrid EDA

申 請 者

CHEN, Benhui

情報生産システム工学専攻
ニューロコンピューティング研究

2010 年 11 月

How to efficiently extract the useful information from high throughput biological data is one of the main challenges in bioinformatics. Machine learning techniques such as Support Vector Machines (SVMs), Neural Networks (NNs) and Estimation of Distribution Algorithms (EDAs) have been effective in analyzing biological data because of their capabilities in handling uncertain data noise and in generalization. Gene function prediction and protein structure prediction are two important domains where machine learning techniques are applied in bioinformatics. From the view of machine learning, the gene function predictions can be seen as multi-label classification or hierarchical multi-label classification (HMC) tasks; the protein structure predictions can be seen as optimization tasks that search for the protein folding structure with minimum energy based on its amino acid sequence.

For complicated tasks in bioinformatics, applying machine learning methods simply usually cannot obtain expectable results. The different characteristics and challenges of the tasks will require novel modifications of existing machine learning methods. In this thesis, some improved methods based on multi-SVM and hybrid EDA are developed for solving complicated tasks in the predictions of gene function and protein HP model. According to the characteristics of applications, the prior knowledge of tasks and the information of biological data are extracted by some delicate techniques. The obtained knowledge and information are used to enhance machine learning methods. SVMs based methods are developed for solving different applications of the gene function prediction. EDAs based methods are proposed for solving protein HP model problems. The Dill's HP-lattice model is a simplified model of protein structure prediction. It is a useful tool for investigating general properties of protein folding.

The thesis is organized from 7 chapters as follows.

Chapter 1 briefly introduces predictions of gene function and protein structure, our motivation, goals of this work and outline of the thesis.

Chapter 2 proposes a new multi-SVM system called composite kernel based SVM (ck-SVM) model for solving nonlinear classification tasks in gene function prediction. In gene function classification, there are some datasets with characteristics of high noise, large number of input features and relatively small number of training examples. For those datasets, conventional nonlinear classification methods are unavailable due to the over-fitting problem. The proposed ck-SVM model realizes a nonlinear separating boundary by estimating a series of

piecewise local linear boundaries. A local linear multi-SVM system is estimated by a composite kernel incorporating prior knowledge of training data. Instead of building multiple local SVM models separately, the prior knowledge of local subsets is used to construct a composite kernel, and the local linear multi-SVM model is realized exactly in the same way as a single SVM model by using the composite kernel. The proposed ck-SVM is a controllable support vector machine model, it can solve nonlinear tasks and avoid over-fitting in certain degree by introducing multiple local linear boundaries.

Chapter 3 proposes a multi-label classification method based on label ranking and delicate decision boundary SVM. In multi-label gene classification, a training example is associated with a set of labels and the task is to predict the label set for each unseen instance. The main challenge of this problem is that classes are usually overlapped and correlated. Label ranking based method requires a real-valued score for each class (label) to order a label rank, then classifies a new instance into the classes that ranking score above a threshold. In order to obtain a proper label rank, an improved probabilistic SVM with delicate decision boundary is proposed as the scoring method. It can improve the probabilistic label rank by introducing the information of overlapped training samples into learning procedure. Instead of estimating a static threshold for all testing instances, an instance-independent thresholding strategy is proposed to decide the classification results. It can estimate an appropriate threshold for each testing instance according to the characteristics of instance and label rank.

Chapter 4 proposes a hierarchical multi-label classification method based on over-sampling and hierarchy constraint. Hierarchical multi-label classification (HMC) is a special variant of multi-label classification where the classes are organized in a hierarchy. Gene FunCat prediction problem is a difficult HMC task. In FunCat datasets, there are hundreds of functional classes structured by a predefined hierarchy, and the example distributions of classes are usually high degree skewed (imbalanced) with negative more than positive. Two measures are implemented in the proposed method to improve the HMC performance by introducing the hierarchy constraint into learning procedures. Firstly, for imbalanced functional classes, a hierarchical SMOTE is proposed as over-sampling preprocessing to enhance the SVM learning efficiency. Secondly, an improved True Path Rule consistency approach is introduced to ensemble the results of binary probabilistic SVM classifications. It can correct the classification results and guarantee the hierarchy constraint of classes.

Chapter 5 proposes a hybrid EDA for solving the protein structure prediction problem on HP model. The protein HP model prediction is an NP-complete optimization problem. The task is to predict the minimum energy lattice structure of a protein from its simplified two-letter (H and P) amino acid sequence. EAs based methods can be used to solve this NP-complete problem. But for long protein sequences, the conventional methods can only find the suboptimum solutions. In order to select better individuals for probabilistic model of EDA, a composite fitness function containing the information of folding structure core (H-Core) is introduced to replace the traditional fitness function of HP model. Then, local search with guided operators is utilized to refine found solutions for improving efficiency of EDA. In addition, an improved backtracking-based repairing method is proposed to repair invalid individuals sampled by the probabilistic model of EDA. It can significantly reduce the number of backtracking searching operation and the computational cost for long sequence protein.

Chapter 6 proposes an adaptive niching EDA with balance searching based on clustering analysis. For optimization problems with irregular and complex multimodal landscapes, EDAs always suffer from the drawback of premature convergence similar to other evolutionary algorithms. In the proposed adaptive niching EDA, the Affinity Propagation (AP) clustering is used to adaptively partition the population into niches during a run of EDA firstly. Then, a niche capacity selection mechanism based on the Boltzmann scheme is introduced to realize a balance searching between exploration and exploitation. Two different selection strategies, novelty-proportionate selection based on the searching history information and fitness-proportionate selection based on the best fitness in a niche, are assembled by a Boltzmann weight. Tuned by the Boltzmann weight, at the beginning phase of searching, the dominating novelty-proportionate selection can guide the EDA searching to cover the search space as much as possible; at the end phase of searching, the dominating fitness-proportionate selection can enforce the EDA searching to exploit the already found promising areas. The proposed adaptive niching EDA is used for solving the protein HP model prediction. In addition, three benchmark functional multimodal optimization problems are also used to evaluate the proposed method.

Chapter 7 summarizes the thesis and gives suggestions for further research.