Data Mining for Building Neural Protein Sequence Classification Systems with Improved Performance

Dianhui Wang Nung Kion Lee Tharam S. Dillon

Department of Computer Science and Computer Engineering
La Trobe University, Melbourne, V.IC 3083, Australia

Abstract - Traditionally, two protein sequences are classified into the same class if their feature patterns have high homology. These feature patterns were originally extracted by sequence alignment algorithms, which measure similarity between an unseen protein sequence and identified protein sequences. Neural network approaches, while reasonably accurate at classification, give no information about the relationship between the unseen case and the classified items that is useful to biologist. In contrast, in this paper we use a generalized radial basis function (GRBF) neural network architecture that generates fuzzy classification rules that could be used for further knowledge discovery. Our proposed techniques were evaluated using protein sequences with ten classes of super-families downloaded from a public domain database, and the results compared favorably with other standard machine learning techniques.

I. INTRODUCTION

A protein super-family consists of protein sequence members that are evolutionally related and therefore functionally and structurally relevant with each other [1,22]. One of the benefits from this category grouping is that some molecular analysis can be carried out within a particular super-family instead of individual protein sequence. It has also become apparent that the function of most genes is still unknown and classification into functionally related groups will provide valuable information on the protein function. Traditionally, two protein sequences are classified into the same class if they have high homology in terms of feature patterns extracted through sequence alignment algorithms. These algorithms, for instance, SAM[11], MEME[12], iPro-Class [9], compare an unseen protein sequence with all the identified protein sequences and provide a score based on similarity of sequences. As the size of the protein sequence databases are large, it is a very time consuming job to perform exhaustive comparison of existing protein sequences. Therefore, it is useful and helpful to build an intelligent classification system for effectively searching protein sequences in some large protein databases. Motivated by this, recently neural networks have been successfully applied in this domain and the results obtained demonstrate some merits of the methodology [1,13]. Neural networks (NNs) have been chosen as technical tools for the protein sequence classification task due to the following two reasons: (i) the extracted features of protein sequences are distributed in a high dimensional space with complex characteristics which is difficult to satisfactorily model using some statistical or parameterized approaches; and (ii) neural networks are able to use the raw continuous values as system inputs. Basically, there are two types of neural models applicable for protein sequences classification task, i.e., unsupervised self-organizing mapping (SOM) networks [8,13] and supervised feed-forward neural networks (FNNs) [14,15,16]. The use of the SOM networks is to discover relationships within a set of protein sequences by clustering them into different groups. In contrast, the FNN based classification systems emphasizes on matching patterns through supervised learning. Once off-line training of the neural network is accomplished, the resulting neural classifier is ready to be used for future protein sequence classification and only few seconds are needed to classify a new protein sequence. This saves a lot of time as compared to sequence alignment methods. Besides the direct protein classification, the supervised neural classifier could also been used to reduce the search scope of the sequence alignment program by only searching members of super-families [22].

Data mining is a process of transferring and analyzing available sets of specific data and extracting the information and knowledge in the form of relationships, patterns or clusters for decision-making, classification, prediction and control [17]. Construction typically involves clustering data points that are close to one another according to some metric or criteria [18]. Given a set of pre-classified examples described in terms of some attributes, the goal of data mining for classification tasks is to derive a set of IF-THEN rules that can be used to assign new events to the appropriate classes. To generate fuzzy classification rules, existing techniques can be categorized into two broad classes: a *direct* method and an *indirect* method. In the *direct* method, the cluster centers as linguistic concepts for fuzzy rules are derived from training data, then the relevant membership functions associated with these cluster centers are assigned by some parameters and further tuned optimally to satisfy some criteria [19]. The *indirect* method encodes domain knowledge expressed using linguistic concepts in various NN models, then updates the structures and weights of the NNs so that the final neural models may classify the given task effectively and efficiently. Note that this method can automatically generate one with explanatory functional fuzzy rules and has a fast inference process due to the connectionist models obtained [4]. For more details, readers may refer to a recent published survey paper [5].

The goal of this paper is to construct a generalized RBF network, which generates a set of fuzzy rules (the direct method), for protein sequence classification tasks. The rest of the paper is organized as follows: Section II discusses some issues on classifier design. Section III presents a novel objective function for classifier optimization. Section IV evaluates the performance of the proposed neural protein sequence classification system, where a data preprocessing description and a comparison are given. We conclude this work with some remarks in the last section.

0-7803-7898-9/03/$17.00 ©2003 IEEE