Available online at www.sciencedirect.com
SCIENCE DIRECT®
Neurocomputing 68 (2005) 315–321
NEUROCOMPUTING
www.elsevier.com/locate/neucom

Letters

Fusion of classifiers for protein fold recognition

Loris Nanni*

DEIS, IEIIT—CNR, Università di Bologna Viale Risorgimento 2, 40136 Bologna, Italy

Received 16 January 2005; received in revised form 1 March 2005; accepted 6 March 2005
Available online 13 May 2005
Communicated by R. W. Newcomb

Abstract

Predicting the three-dimensional structure of a protein from its amino acid sequence is an important problem in bioinformatics and a challenging task for machine learning algorithms. Given (numerical) features, one of the existing machine learning techniques can be then applied to learn and classify proteins represented by these features. We show that combining Fisher's linear classifier and *K*-Local Hyperplane Distance Nearest Neighbor we obtain an error rate lower than previously published in the literature.

Keyword: Fusion of classifiers; Protein fold recognition; Machine learning algorithms

1. Introduction

The folding problem is central in molecular biology and it can be formulated as follows: given the primary structure (linear sequence of amino acids in a protein molecule) of a protein, how the three-dimensional (3-D) fold can be deduced from it. To be able to predict the protein structure from the amino-acid sequence would have tremendous impact in all of biotechnology and drug design.

Recently, several works have approached the problem of predicting the 3-D structure of a protein by applying techniques from machine learning [3] employed ensembles of both three-layer feed-forward neural networks (NN) (trained with the

*Corresponding author.
E-mail address: lnanni@deis.unibo.it.

0925-2312/$ - see front matter
doi:10.1016/j.neucom.2005.03.001

conjugate gradient method) and support vector machines (one-vs.-all, unique one-vs.-all and one-vs.-one methods for building multi-class support vector machines). Each ensemble consisted of many two-class classifiers. [1] used a 131-dimensional feature vector and an ensemble of four-layer discretized interpretable multi-layer perceptrons (DIMLP), where each network learns all protein folds simultaneously. Bagging and arcing combined the outputs of DIMLPs. [2] selected NN and support vector machines (SVM) as basic building blocks of two-level classification. The NN with a single hidden layer were used: multi-layer perceptron (MLP), radial basis function network (RBFN), and general regression neural network (GRNN). [9] performed independent classification at each level, i.e. they did not utilize a hierarchical classifier. They trained the MLP and RBFN with new features (400 in number) based on the hydrophobicity of the amino acids. In [7] a modified nearest neighbor algorithm called the *K*-Local Hyperplane is applied to protein fold recognition based on features derived from secondary structure.

The combination of multiple classifiers was shown to be suitable for improving the recognition performance in difficult classification problems [6].

We show that combining Fisher's linear classifier and *K*-Local hyperplane distance nearest neighbor, we obtain an error rate lower than previously published in the literature.

2. System

To use machine learning methods, feature vectors are extracted from protein sequences. Six features were extracted from protein sequences: amino acids' composition, predicted secondary structure, hydrophobicity, normalized van der Waals volume, polarity, and polarizability. All (except the first) features have dimensionality 21, whereas the composition has dimensionality 20. Thus, in total, a feature vector combining six features has 125 dimensions (or components). In addition, the protein length is reported. As a result, in total, a feature vector has the length of 126. *K*-local hyperplane distance nearest neighbor (HKNN) [7] is trained using these 126 features, these features are projected onto a 126-dimensional space by Karhunen Loeve transform (KL) [4]. Fisher's linear classifier is trained using the features obtained by KL Fig. 1.

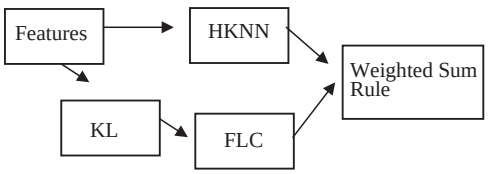

Fig. 1. System proposed.