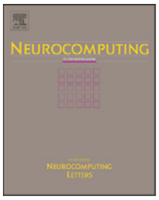
Neurocomputing 75 (2012) 194–198

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



A hybrid discriminative/generative approach to protein fold recognition

Wiesław Chmielnicki [a,*], Katarzyna Stąpor [b]

[a] Jagiellonian University, Faculty of Physics, Astronomy and Applied Computer Science, ul. Reymonta 4, 30-059 Kraków, Poland

[b] Silesian University of Technology, Institute of Computer Science, ul. Akademicka 16, 44-100 Gliwice, Poland

ARTICLE INFO

Available online 30 July 2011

Keywords:
Support vector machine
Statistical classifiers
RDA classifier
Protein fold recognition

ABSTRACT

There are two standard approaches to the classification task: generative, which use training data to estimate a probability model for each class, and discriminative, which try to construct flexible decision boundaries between the classes. An ideal classifier should combine these two approaches. In this paper a classifier combining the well-known support vector machine (SVM) classifier with regularized discriminant analysis (RDA) classifier is presented. The hybrid classifier is used for protein structure prediction which is one of the most important goals pursued by bioinformatics. The obtained results are promising, the hybrid classifier achieves better result than the SVM or RDA classifiers alone. The proposed method achieves higher recognition ratio than other methods described in the literature.

1. Introduction

Protein structure prediction is one of the most important goals pursued by bioinformatics. The structure of a protein (fold) is closely related to its biological function [1] so it is very important to know not only sequence of amino acids in a protein molecule, but also how this sequence is folded. The completion of many genome-sequencing projects has meant that the number of proteins with known amino acids sequence is quickly increasing, but the number of proteins with known 3D structure is still relatively very small.

There are several machine-learning methods to predict the protein folds from amino acid sequences proposed in the literature. Ding and Dubchak [2] experimented with support vector machine (SVM) and neural network (NN) classifiers. Shen and Chou [3] proposed ensemble model based on nearest neighbour. A modified nearest neighbour algorithm called K-local hyperplane (HKNN) was used by Okun [4]. Nanni [5] proposed ensemble of classifiers: Fishers linear classifier and HKNN classifier.

Another group of methods commonly used are profile hidden Markov models (HMMs) [6]. They are amongst the most successful procedures for detecting remote homology between proteins. There are two popular profile HMM programs, HMMER and SAM. The comparison between them can be found in [7]. However, the main drawback of HMMs is the employment of large model architectures which require large data sets and high computational effort for training. This problem was partially solved by introducing a reduced state-space HMM with a much smaller architecture, see [8,9].

There are two standard approaches to the classification task: generative classifiers use training data to estimate a probability model for each class, then test items are classified by comparing their probabilities under these models. The discriminative classifiers try to find the optimal frontiers between classes based on all samples of the training data set. This paper presents a classifier which combines the SVM (discriminative) classifier with statistical RDA (generative) classifier.

The fusion of the different classifiers is widely used in Bioinformatics to improve the performance. For example Shen and Chou [10] proposed an ensemble classifier for large-scale human protein subcellular location prediction or Nanni et al. [11] presented series of SVM classifiers combined with the max rule for two problems: HIV-protease and recognition of T-cell epitopes. An interesting method of ensemble classifier generation called RotBoost is described in [12].

The SVM is a binary classifier but the protein fold recognition is a multi-class problem. There are many methods proposed to deal with this issue. One of the first and well-known methods is one-versus-one strategy with max-win voting scheme [13]. In this strategy every binary classifier votes for the preferred class and the voting table is created. Originally a class with the maximum number of votes is recognized as the correct class.

However, some of these binary classifiers are unreliable. The votes from these classifiers influence the final classification result. In this paper there is a strategy presented to assign a weight to each vote based on the values of the discriminant function from RDA classifier.

The rest of this paper is organized as follows: Section 2 introduces the database and the feature vectors used is the

[*] Corresponding author. Tel.: +48 605 689692.
E-mail address: wieslaw.chmielnicki@uj.edu.pl (W. Chmielnicki).

0925-2312/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2011.04.033

experiments, Section 3 describes the method of combining the classifiers, Section 4 presents experimental settings and results and Section 5 offers the conclusions as well as the future work.

2. The database and feature vectors

Using machine-learning methods entails the necessity to find out databases with representation of known protein sequences and its folds. Then this information must be converted to the feature space representation.

2.1. Database

In experiments described in this paper two data sets derived from the structural classification of proteins (SCOP) database [14] are used. The detailed description of these sets can be found in [2]. The training set consists of 313 protein sequences and the testing set consists of 385 protein sequences. These data sets include proteins from 27 most populated different classes (protein folds) representing all major structural classes: α, β, α/β, and $\alpha+\beta$. The training set was based on PDB_select sets [15,16] where two proteins have no more than 35% of the sequence identity. The testing set was based on PDB-40D set [17] from which representatives of the same 27 largest folds are selected. The proteins that had higher than 35% identity with the proteins of the training set are removed from the testing set.

2.2. Feature vectors

In our experiments the feature vectors developed by Ding and Dubchak [2] were used. These feature vectors are based on six parameters: Amino acids composition (C), Predicted secondary structure (S), Hydrophobity (H), Normalized van der Waals volume (V), Polarity (P) and Polarizability (Z). Each parameter corresponds to 21 features except Amino acids composition (C), which corresponds to 20 features. The data sets including these feature vectors are available at http://ranger.uta.edu/~chqding/protein/. For more concrete details, see [18,19].

The feature vector was slightly changed. The length of the amino acid sequence was added to the Amino acids composition (C) vector, so now the C vector has also $20+1=21$ features. Therefore the full feature vector (C, S, H, V, P, Z) counts $6\times 21=126$ features. All values of the feature vectors are scaled to the range $[-1;+1]$ before applying an SVM classifier. The main advantage of scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges.

3. The proposed combined classifier

The discriminative classifiers are based on minimum error training, for which the parameters of one class are trained on the samples of all classes. For statistical classifiers, the parameters of one class are estimated from the samples of its own class only. Therefore the characteristics of these kinds of classifiers differs in several respects.

The discriminative classifiers give higher accuracies than statistical ones when there is enough training samples, but however the accuracy of regularized statistical classifiers (such as RDA) are more stable and when training data set is small they generalize better.

Additionally, the statistical classifiers are resistant to outliers, whereas the discriminative ones are susceptible to outliers because their decision regions tend to be open [20]. For more detailed discussion see [21].

In protein fold recognition problem we have very small training data sets. So our motivation was to combine the properties of both types of classifiers.

3.1. The SVM classifier

The support vector machine (SVM) is a well-known large margin classifier proposed by Vapnik [22]. The basic concept behind the SVM classifier is to find an optimal separating hyperplane, which separates two classes. The decision function of the binary SVM is

$$f(x)=sign\left(\sum_{i=1}^{N}\alpha_i y_i K(x_i,x)+b\right), \tag{1}$$

where b is a constant, $y_i\in\{-1,1\}$, $0\le\alpha_i\le C, i=1,2,\ldots,N$ are nonnegative Lagrange multipliers, C is a cost parameter, that controls the trade-off between allowing training errors and forcing rigid margins, x_i are the support vectors and $K(x_i,x)$ is the kernel function.

The SVM is a binary classifier but the protein fold recognition is a multi-class problem. There are many methods proposed in the literature to deal with this issue, such as one-versus-others, one-versus-one strategies, DAG (directed acyclic graph), ADAG (ADAPTIVE DIRECTED ACYCLIC GRAPH) methods [23,24], BDT (binary decision tree) approach [25], DB2 method [26], pairwise coupling [27] or error-correcting output codes [28]. In our experiments we use one-versus-one strategy with max-win voting scheme.

3.2. The RDA classifier

Quadratic discriminant analysis (QDA) [29] models the likelihood of a class as a Gaussian distribution and then uses the posterior distributions to estimate the class for a given test vector. This approach leads to the discriminant function:

$$d_k(x)=(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)+\log|\Sigma_k|-2\log p(k), \tag{2}$$

where x is the test vector, μ_k is the mean vector, Σ_k is the covariance matrix and $p(k)$ is the prior probability of the class k. The Gaussian parameters for each class can be estimated from the training data set, so the values of Σ_k and μ_k are replaced in formula (2) by its estimates $\hat{\Sigma}_k$ and $\hat{\mu}_k$.

However, when the number of training samples is small, compared to the number of dimensions of the training vector, the covariance estimation can be ill-posed. The approach to resolve the ill-posed estimation is to regularize the covariance matrix Σ_k. It can be replaced by the average matrix i.e. $\hat{\Sigma}=\sum\hat{\Sigma}_k/\sum\hat{N}_k$ which leads to linear discriminant analysis (LDA). This assumes that all covariance matrices are similar. It is a very limited approach. In regularized discriminant analysis (RDA) [30] each covariance matrix is estimated as

$$\hat{\Sigma}_k(\lambda)=(1-\lambda)\hat{\Sigma}_k+\lambda\hat{\Sigma}, \tag{3}$$

where $0\le\lambda\le 1$. The parameter λ controls the degree of shrinkage of the individual class covariance matrix estimate toward the average estimate.

3.2.1. Feature selection

As we stated in the previous section our problem is ill-posed. Additionally when the number of the samples is very small, the regularization may be insufficient to solve the problem [31]. In our experiments we used the selection algorithms to reduce dimensionality of the feature space.

There are many feature selection algorithms described in the literature. However, we look for a simple, but effective approach