Neurocomputing 71 (2008) 3104–3113

Contents lists available at ScienceDirect

Neurocomputing

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

Fusion of feature selection methods for pairwise scoring SVM

Man-Wai Mak[a,*], Sun-Yuan Kung[b]

[a] Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong SAR, Hong Kong
[b] Department of Electrical Engineering, Princeton University, USA

ARTICLE INFO

Available online 21 June 2008

Keywords:
Feature selection
Filter
Wrapper
Curse of dimensionality
Protein sequences
Subcellular localization
Kernel methods
Support vector machines

ABSTRACT

It has been recently discovered that stacking the pairwise comparison scores between unknown patterns and a set of known patterns can result in feature vectors with desirable discriminative properties for classification. However, such technique can be hampered by the curse of dimensionality because the vectors size is equal to the training set size. To overcome this problem, this paper investigates various filter and wrapper feature selection techniques for reducing the feature dimension of pairwise scoring matrices and argues that these two types of selection techniques are complementary to each other. Two fusion strategies are then proposed to (1) combine the ranking criteria of filter and wrapper methods at algorithmic level and (2) merge the features selected by the filter and wrapper methods. Evaluations on a subcellular localization benchmark and a microarray dataset demonstrate that feature subsets selected by the fusion methods are either superior to or at least as good as those selected by the individual methods alone for a wide range of feature dimensions.

1. Introduction

In computational biology, the subcellular location and structural family of a protein provide important information about its biochemical functions. However, experimental analysis of proteins is time-consuming and cannot be performed on genome-wide scales. Therefore, a reliable and efficient method is essential for automating the prediction of proteins' subcellular locations and the classification of protein sequences into functional and structural families. One of the successful techniques is to compare the unknown sequences against some known sequences. The idea is based on the notion that similarity (homology) in sequences, to a certain extent, also means closeness in functions and structures.

The comparison of two sequences are often hampered by the fact that the two sequences often have different lengths whether or not they belong to the same family. To overcome this problem, performing pairwise comparisons between a sequence and a set of known sequences has been a popular approach to creating fixed-size feature vectors from variable-length sequences [13,7,5]. Although this pairwise approach can usually create feature vectors with good discriminative properties, it also has its own limitation. The main problem is that the feature dimension is the same as the number of training patterns. This leads to the curse of dimensionality, because the training set size could be very large. In fact, for the applications addressed in this paper, they are in the range of several thousands. High dimensionality in feature space increases the computational cost in both the learning phase and prediction phase. In the prediction phase, the more features used the more the computation required and the lower the retrieval speed. Fortunately, the prediction time is often linearly proportional to the number of features selected. Unfortunately, in the learning phase, the computational demand may grow exponentially with the number of features. Because a large number of sequences are being added to sequence databases in a daily basis, it is imperative to reduce the complexity of the pairwise scoring approach.

An obvious solution to the curse of dimensionality problem is to reduce the feature size and yet retaining the most important information critical for the classification of the training patterns. Research in protein homology detection has found that just over 10% of proteins' profiles contribute 90% of the total score for positive training sequences [11], suggesting that some features are more relevant to the classification task than the others. The feature size can be reduced by either finding principle subspace or weeding out those less significant features. The latter approach is known as feature selection and is the focus of this paper.

Feature selection often depends on a joint consideration of two conflicting aspects: computational cost and achievable performance. The best choice usually represents an optimal tradeoff between these factors. The challenge lies in how to reach a useful dimension reduction while conceding minimum sacrifice on accuracy or other desired performance.

[*] Corresponding author.
E-mail address: enmwmak@polyu.edu.hk (M.-W. Mak).
URL: http://www.eie.polyu.edu.hk/~mwmak/mypage.htm (M.-W. Mak).

0925-2312/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.neucom.2008.04.024

Feature selection methods can be divided into two categories: filter and wrapper. In the filter method, feature selection and classifier design are separated in that a subset of features is firstly selected and then classifiers are trained based on the selected features. For example, the discriminative power of features are ranked according to their Fisher discriminant ratio (FDR) [3] such that features with high ranks are retained. The wrapper approach [24,6,25,12], on the other hand, uses classification accuracies or criteria derived from the classifier to rank the discriminative power of all (or part) of the possible feature subsets and select the subset that produces the best performance. Therefore, the selected features are bound to the type of classifier that was used in the feature selection process.

Both filter and wrapper methods have their own merits and limitations. For example, although filter methods are simple and fast, the selected features may be highly correlated because the feature set may contain many highly discriminative features but with almost identical characteristics. Moreover, most ranking criteria do not take the combined effect of features into account. This paper proposes fusing the filter and wrapper methods to leverage the advantages of both methods. Evaluations on a subcellular localization task and a gene selection problem demonstrate that the two feature selection approaches are complementary to each other.

The paper is organized as follows. Section 2 outlines the profile alignment algorithms and the pairwise scoring kernels for protein subcellular localization. Section 3 describes two important feature selection strategies: filter and wrapper, and Section 4 proposes two fusion techniques to combine both strategies. The fusion methods are then compared with some existing selection methods through a subcellular localization task and a gene selection task in Section 5, and conclusions are drawn in Section 6.

2. Pairwise scoring kernels for subcellular localization

Denote $\mathscr{D} = \{S^{(1)}, \ldots, S^{(N)}\}$ as a training set containing N protein sequences. To efficiently produce the profiles of a protein sequence (called query sequence), the sequence is used as a seed to search and align homologous sequences from protein databases such as Swissprot [26] using the PSI-BLAST program [1] as shown in Fig. 1. Let us denote the operation of PSI-BLAST search given the query sequence $S^{(i)}$ as

$$\phi^{(i)} \equiv \phi(S^{(i)}) : S^{(i)} \longrightarrow \{\mathbf{P}^{(i)}, \mathbf{Q}^{(i)}\},$$

where $\mathbf{P}^{(i)}$ and $\mathbf{Q}^{(i)}$ are the position-specific scoring matrix (PSSM) and position-specific frequency matrix (PSFM) of $S^{(i)}$, respectively. The homolog information pertaining to the aligned sequences is represented by these two matrices (also called profiles), which have 20 rows and n_i columns, where n_i is the number of amino acids in the query sequence. Because these matrices contain rich information about the remote homolog of the query sequence, highly discriminative features can be derived from them for the prediction of subcellular locations and protein functions.

Given the profiles of two sequences $S^{(i)}$ and $S^{(j)}$, we can apply the Smith–Waterman algorithm [20] and its affine gap extension [4] to align $\mathbf{P}^{(i)}$, $\mathbf{Q}^{(i)}$, $\mathbf{P}^{(j)}$, and $\mathbf{Q}^{(j)}$ to obtain the normalized profile alignment score $\zeta(\phi^{(i)}, \phi^{(j)})$, as detailed in the Appendix. The scores $\{\zeta(\phi^{(i)}, \phi^{(j)})\}_{i,j=1}^{N}$ constitute a symmetric matrix $\mathbf{Z}$ whose columns can be considered as N-dimensional vectors:

$$\zeta^{(j)} = [\zeta(\phi^{(1)}, \phi^{(j)}) \ \ldots \ \zeta(\phi^{(N)}, \phi^{(j)})]^{\mathsf{T}}, \quad j = 1, \ldots, N. \tag{1}$$

This means that there are N feature vectors with dimension equal to the training set size. The N N-dimensional column vectors can be used to train M one-versus-rest SVMs for an M-class protein

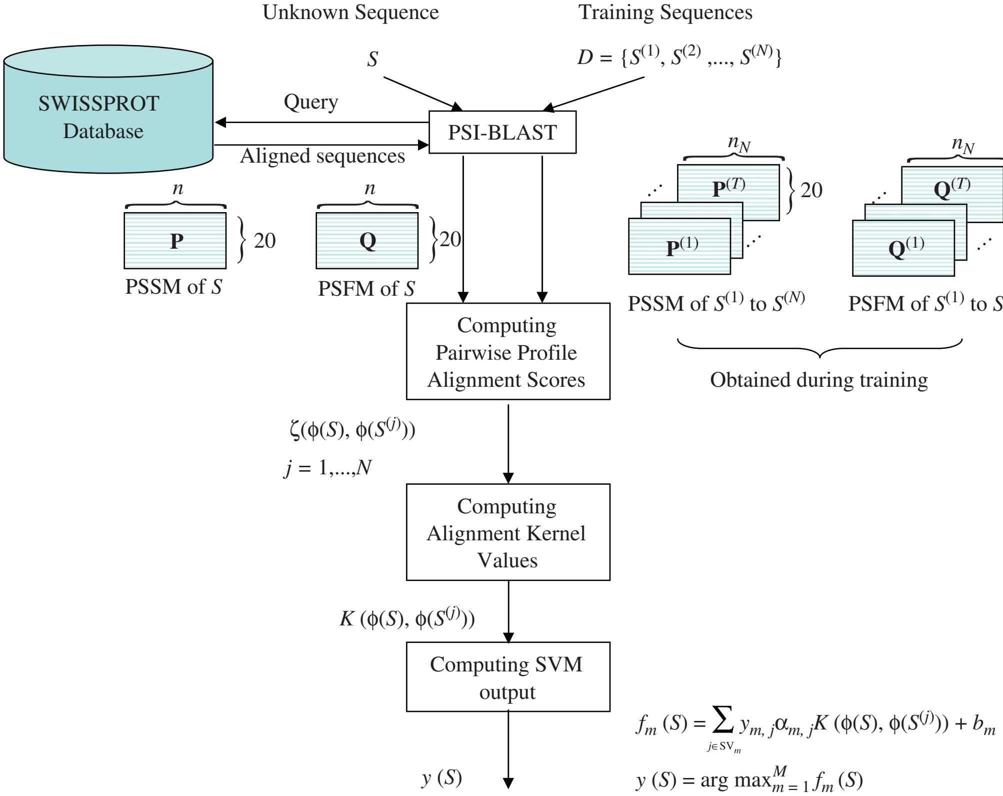

Fig. 1. Prediction of subcellular locations of proteins by pairwise profile alignment SVMs.