Contents lists available at ScienceDirect

Computational Biology and Chemistry

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

Enzyme classification using multiclass support vector machine and feature subset selection

Debasmita Pradhan[a,*], Sudarsan Padhy[a], Biswajit Sahoo[b]

[a] *Department of Computer Scienceing and Engineering, Silicon Institute of Technology, Silicon Hills, Patia, Bhubaneswar, 751024, India*
[b] *School of Computer Engineering, KIIT University, Bhubaneswar, 751024, India*

ARTICLE INFO

Article history:
Received 25 May 2017
Received in revised form 15 July 2017
Accepted 15 August 2017
Available online 31 August 2017

Keywords:
Enzyme classification
Multiclass SVM
Sequential Forward Floating Selection (SFFS)
Orthogonal Forward Selection (OFS)
SVM Recursive Feature Elimination (SVM-RFE)
Random Forest (RF)

ABSTRACT

Proteins are the macromolecules responsible for almost all biological processes in a cell. With the availability of large number of protein sequences from different sequencing projects, the challenge with the scientist is to characterize their functions. As the wet lab methods are time consuming and expensive, many computational methods such as FASTA, PSI-BLAST, DNA microarray clustering, and Nearest Neighborhood classification on protein–protein interaction network have been proposed. Support vector machine is one such method that has been used successfully for several problems such as protein fold recognition, protein structure prediction etc. Cai et al. in 2003 have used SVM for classifying proteins into different functional classes and to predict their function. They used the physico-chemical properties of proteins to represent the protein sequences. In this paper a model comprising of feature subset selection followed by multiclass Support Vector Machine is proposed to determine the functional class of a newly generated protein sequence. To train and test the model for its performance, 32 physico-chemical properties of enzymes from 6 enzyme classes are considered. To determine the features that contribute significantly for functional classification, Sequential Forward Floating Selection (SFFS), Orthogonal Forward Selection (OFS), and SVM Recursive Feature Elimination (SVM-RFE) algorithms are used and it is observed that out of 32 properties considered initially, only 20 features are sufficient to classify the proteins into its functional classes with an accuracy ranging from 91% to 94%. On comparison it is seen that, OFS followed by SVM performs better than other methods. Our model generalizes the existing model to include multiclass classification and to identify most significant features affecting the protein function.

1. Introduction

Proteins are important macromolecules responsible for almost all biological processes in a cell such as growth, function, cell metabolism and maintenance. With the availability of large no of biological sequences obtained from different sequencing projects (Koonin et al., 1998a; Fetrow and Skolnick, 1998), the challenge with the scientist is to know the functions of the newly generated protein sequences in order to understand the biological processes (Siomi and Dreyfuss, 1997; Draper, 1999; Koonin et al., 1998b).

There are many methods available for functional annotation of newly sequenced proteins. The wet lab method of functional characterization of proteins is time consuming and expensive, where as computational approaches are fast and cost effective. The classical computational approaches for function prediction use programs like FASTA (Pearson and Lipman, 1988) and PSI-BLAST (Altschul et al., 1990) which are based on homology between the annotated sequences with unannotated sequence i.e the new sequence. The methods of Comparative Genomics are also used for the prediction of protein function (Pellegrini et al., 1999). They consider the protein to be functionally linked if they have similar phylogenetic profiles (Marcotte et al., 2000; Zheng et al., 2002). Some authors such as David J. Lockhart et al., Mark Schena (Lockhart et al., 1996; Schena et al., 1995), designed clustering algorithms to be used on DNA-microarray data to predict the protein function based on the assumption that genes with correlated expression profile are functionally related (Eisen et al., 1998; Zhou et al., 2002). The protein-protein-interaction

[*] Corresponding author.
E-mail address: debasmita@silicon.ac.in (D. Pradhan).

http://dx.doi.org/10.1016/j.compbiolchem.2017.08.009
1476-9271/© 2017 Elsevier Ltd. All rights reserved.

networks are also used for prediction of protein function using Nearest Neighborhood approach (Lin et al., 2006) based on the fact that proteins may interact for a common purpose. But as protein-protein-interaction data is noisy, the prediction accuracy becomes low. Some methods (Tatusov et al., 2001; Jones et al., 2014) predict the function of a protein by classifying it into a specific functional class based on the sequence similarity. These methods work well if the similarity between sequences is significant. However the prediction becomes random if the similarity between two sequences is not up to a threshold.

Support vector machine method (Vapnik, 2013) is used for protein fold recognition (Ding and Dubchak, 2001; Cai et al., 2002a), protein structure prediction (Yuan et al., 2002; Hua and Sun, 2001; Cai et al., 2002b), protein–protein interaction prediction, and protein function classification (Cai et al., 2003). In these problems the physico-chemical properties of proteins computed from sequences, are used as input for implementing the method. Cai et al. (Cai et al., 2003) used Binary SVM classifier to predict the functional class of a protein. They considered the functional classes like RNA-binding proteins, protein homodimers, drug absorption proteins, drug delivery proteins, drug excretion proteins, Class-I drug metabolizing enzymes, Class-II drug metabolizing enzymes and used 1808 physico-chemical properties such as hydrophobicity, polarity, polarizability, charge, surface tension, secondary structure etc. to represent a protein sequence and obtained accuracy in the range 88%–99% for different classes. Moreover as the dimension of the feature vector used is very high the computation takes more time.

In our model at the first step a binary classifier is designed to classify a protein sequence as enzyme or non-enzyme. In the second step a multi-class classifier is designed to predict the functional class of the protein out of six available enzyme classes such as oxidoreductases, transferases, hydrolases, lyases, isomerase, and ligases. To implement the model, initially 32 physico-chemical properties like number of amino acids, theoretical pie, amino acid compositions(20), number of negatively charged residue, number of positively charged residue, atomic compositions(5), aliphatic index, and hydrophobicity are considered. Since many of the features may carry redundant information, Sequential Forward Floating Selection algorithm (SFFS) (Pudil et al., 1994), Orthogonal Forward Selection (OFS) (Mao, 2004) algorithm, and SVM Recursive Feature Elemination(SVM-RFE) (Guyon et al., 2002; Rakotomamonjy, 2003) are applied to identify the most significant features for classifying the proteins. SFFS gives amino acid compositions such as Arg(A), Asn(N), Cys(C), Gln(Q), Glu(E), Ile (I), Leu(L), Lys(K), Met(M), Phe(F), Pro(P), Ser(S), Thr(T), Trp(W), Tyr (Y), Val(V), atomic compositions, such as Hydrogen(H), Nitrogen (N), Oxygen(O), Sulfur(S) are more significant features where as OFS gives aliphatic index, number of amino acids, atomic compositions such as Carbon(C), Oxygen(O), amino acid compositions such as Cys(C), Asp(D), Arg(R), Phe(F), Gly(G), Pro(P), His(H), Ile(I), Thr(T), Trp(W), Leu(L), Gln(Q), Lys(K), Try(Y), no of positively charge residues, and no of negatively charged residues are more significant features. However, when SVM-RFE is applied it dropped seven features such as number of amino acids, Theoritical pie, Cys (C), Gly(G), Ile(I), Carbon(C), Sulfur(S) to yield 25 significant features and with these features an accuracy range of 90.6149%–93.5275% is obtained. Results of these three algorithms show that Gln(Q), Leu(L), Lys(K), Phe(F), Pro(P), Thr(T), Trp(W), Tyr(Y), and Oxygen(O) play major role for functional classification of proteins. Using all 32 features, i.e Without Feature Selection(WFS) an accuracy range from 90.9699% to 93.6455% is obtained where as using Sequential Forward Feature Floating Selection (SFFS) algorithm with 20 significant features an accuracy from 90.3010% to 92.3077% is obtained and using Orthogonal Forward Feature Selection algorithm (OFS) with 20 significant features an accuracy from 89.6321% to 94.3144% is obtained. Our model found that 20 (Atomic and Amino acid compositions) out of 32 physico-chemical properties are sufficient to predict the functional class of a protein with a high accuracy. The performance of our model is compared with the Random Forest classification algorithm (Liaw and Wiener, 2002). The average accuracy obtained by Random Forest Model is 86.7314%. It is observed that all the three models discussed above have better average accuracy than Random Forest Model.

The rest of the paper is organized as follows. Section 2 presents Multiclass Support Vector Machine, Sequential Forward Feature Selection algorithm, and Orthogonal Forward Feature Section algorithm. Section 3 describes the proposed model. Section 4 discusses the result and performance of our model and Section 5 concludes the work.

2. Preliminaries

2.1. Multiclass support vector machine

The Support vector machine described in appendixA is a binary classifier i.e it classifies objects belonging to two distinct classes. However the real world problems deal with classifying objects into more than two classes. There are many approaches followed to use SVM for multiclass classification. Following are the frequently used approaches.

2.1.1. One verses the rest classification

This approach constructs as many support vector machines as there are classes in the classification problem i.e. given M classes it constructs M binary SVM classifiers $f_1, \ldots, f_M$. To construct f_i, the i^{th} classifier $(i = 1, \ldots, M)$ an SVM is designed by considering the patterns of i^{th} class as positive samples and patterns of the rest of classes as negative samples. An unknown sample X is classified by providing it to each classifier and applying majority voting technique. The class lebel with maximum frequency is assigned to the pattern X. One of the major limitation of this approach is the training samples used to build the model are highly unbalanced.

2.1.2. Pairwise classification

The pair wise classification technique avoids the limitation of the above method by constructing decision surfaces for each pair of classes. Given the training set $D = \{(x_i, y_i)\}$, $x_i \in R^n$ and $y_i \in \{1, 2, 3 \ldots, M\}$ this method generates $M(M-1)/2$ classifiers, one classifiers for each pair of classes. Let f_{ij} be the classifier which separates the pair of classes i and j with $i \neq j$ and $i, j \in \{1, 2, \ldots, M\}$. $f_{ij}$is trained taking D_i as the positive class and D_j as the negative class where D_i is the samples in D with class level i .The output of the classifier f_{ji} is $-f_{ij}$.once the classifiers are trained an unknown sample X is classified by presenting it to each of $M(M-1)/2$classifiers. Each classifier assigns a class lebel to the new sample. The class lebel with highest count is then considered as the label of the unknown sample X.

2.2. Feature subset selection

Given a set of n features the goal of feature Subset Selection is to select a subset of d features ($d < \mathrm{n}$) without significantly degrading the performance of the recognition system (Pudil et al., 1994).

2.2.1. Sequential Forward Floating Selection Method (SFFS) (Pudil et al., 1994)

This method start with an empty feature set to begin with. In successive steps the features are included/excluded depending on some class separability measure. We use class separability