A Novel Algorithm for Predicting β-barrel Outer Membrane Proteins Using ACO-based Hyper-parameter Selection for LS-SVMs

Guang-ming Xian
Information Engineering and Technology Department,
South China Normal University,
Guangdong Foshan, China, 528225

Biqing-Zeng Xian
Information Engineering and Technology Department,
South China Normal University,
Guangdong Foshan, China, 528225

***Abstract*—An ACO-based hyper-parameter selection for least squares support vector machines (LS-SVMs) was trained to predict the topology of transmembrane β strands proteins. It should be stressed that it is very important to do a careful model selection of the tuning parameters for LS-SVM. In this paper, a novel hyper-parameter selection method for LS-SVMs is presented based on the ant clony optimization (ACO). Optimal LS-SVMs parameters for RBF kernel are selected to predict the topology of the transmembrane β strands proteins. The feasibility of this method is examined on one test database set. For the testing database, the present LS-SVMs method with RBF kernel predicts higher accuracy than SVM and HMM method. The simulation result shows that this prediction model for transmembrane β strands proteins is accurate.**

Keywords-transmembrane β strands; LSSVM; ACO; prediction of membrane protein; parameters Optimization

I. Introduction

Membrane proteins are an abundant and functionally relevant subset of proteins that putatively include from about 15 up to 30% of the proteome of organisms fully sequenced. These estimates are mainly computed on the basis of sequence comparison and membrane protein prediction. It is therefore urgent to develop methods capable of selecting membrane proteins especially in the case of outer membrane proteins, barely taken into consideration when proteome wide analysis is performed. This will also help protein annotation when no homologous sequence is found in the database [1].

At present, two types of membrane proteins have been characterized: The first includes all the proteins that to a different extent interact with the lipid bilayer of the cytosplamic membrane of all cells [2]; the second group includes those proteins that during the last 10 year have been discovered in the outer membrane of bacteria, chloroplasts, and mitochondria [3]. A major distinguishing feature of membrane proteins of the first type is that they span the cytoplasmic membrane with α-helixes, whereas those of the second type interact with the outer membrane with antiparallel β-strands forming barrels, existing as monomers and oligomers [4]. These chains, referred to as β-barrel membrane proteins [3,5], comprise the archetypal trimeric porins of Gram-negative bacteria consisting of water-filled channels that nonspecifically mediate the passive transport of ions and small hydrophilic molecules (<6 kD) or select for certain molecules such as malto-oligosaccha-rides [6].

Outer membrane proteins solved so far at atomic resolution interact with the external membrane of bacteria with a characteristic β barrel structure comprising different even numbers of β strands (β barrel membrane proteins). In this they differ from the membrane proteins of the cytoplasmic membrane endowed with α helix bundles (all alpha membrane proteins) and need specialized predictors [1].

Support vector machine (SVM) has been applied to predict the transmembrane β-strands of the outer membrane proteins. Paper [7] describes a method developed for predicting transmembrane β-barrel regions in membrane proteins using machine learning techniques: artificial neural network (ANN) and SVM. The SVM model was modified by adding 36 physicochemical parameters to the amino acid sequence information. ANN- and SVM-based methods were combined to utilize the full potential of both techniques. In paper [8], three feature classes were calculated from protein sequences: amino acid compositions, dipeptide compositions and weighted amino acid index correlation coefficients. Then, three feature classes were combined and inputted into a support vector machine (SVM) based predictor to identify Outer membrane proteins (OMPs) from other folding types of proteins.

According to Vapnik's "the nature of statistical learning theory" [9], using tactics such as introducing a kernel function, both nonlinear pattern recognition problems and regression problems can be converted into linear ones, and finally deduced to mathematical problems of Quadratics Programming (QP). This category of SVM uses the inherently sparse loss functions, such as epsilon-insensitive loss function, Laplacian loss function, Huber's robust loss function and so on [10]. They are derived from statistical tools and theories, leading to sparse and robust approximations of certain problems [11-13]. However, it requires solving a QP with inequality constraints, which is complicated and time consuming. And to keep the sparseness and robustness of estimation, loss function should be carefully chosen depending on the problem [14].

Least squares support vector machines (LS-SVMs) are introduced by Suykens et al. as reformulations to standard SVMs [15-16] which simplify the training process of standard SVM in a great extent by replacing the inequality constraints with equality ones. The simplicity of LS-SVMs promotes the applications of SVM and many pattern recognition and regression problems have been tackled with LS-SVMs in the last decade [17].

Similar to SVM, LS-SVMs also has the problem of parameter selection. A novel algorithm of parameter selection is proposed based on the principles of the ant colony optimization (ACO). ACO is based on the behavior of a group of artificial ants in search of a shortest path from the source to the destination. These artificial ants mimic real ants in nature in search of food from the nest to the destination. The ants deposit a chemical substance called pheromone that other ants can sense on their journey to the destination. The ants interact with each other and the environment using the pheromone concentration. As with any perfume, if not reapplied, the scent evaporates. As the ants travel, the longer paths lose their pheromone concentration making the ants to choose the shortest path [18]. In this paper, ACO is applied to select the hyper-parameters of LS-SVMs.

II. Least squares support vector machines for nonlinear function estimation

For a regression problem with training set $\{x_k, y_k\}$, $k = 1, 2, \cdots, N$, $x_k = R^n, y_k = R$, in the primal weight space, optimal problems of LS-SVMs can be formulated as follows:

$$\min_{w,b,e} J(w,e) = \frac{1}{2} w^T w + \frac{1}{2}\gamma \sum_{i=1}^{N} e_i^2, \quad \gamma > 0 \tag{1}$$

Subject to $y_i = w^T \varphi(x_i) + b + e_i, i = 1, \cdots, N$ (2)

where $\varphi(\cdot): R^n \to R^{n_h}$ is a function mapping the input space into a so-called higher dimensional (possibly infinite dimensional) feature space, weight vector $w \in R^{n_h}$ is in primal weight space, $e_i \in R$ is error variables, b is bias term and γ is an adjustable constant.

According to Eq. (1), we construct the Lagrangian function as

$$L(w,b,e,\alpha) = J(w,e) - \sum_{i=1}^{N} \alpha_i \left\{ w^T \varphi(x_i) + b + e_i - y_i \right\} \tag{3}$$

where $\alpha_i (i = 1, \ldots N)$ are the Lagrange multipliers (called support vector). The conditions for optimum are given by

$$\begin{cases} \dfrac{\partial L}{\partial w} = 0 \to w = \sum_{i=1}^{N} \alpha_i \varphi(x_i) \\ \dfrac{\partial L}{\partial b} = 0 \to \sum_{i=1}^{N} \alpha_i = 0 \\ \dfrac{\partial L}{\partial e_i} = 0 \to \alpha_i = \gamma e_i, i = 1, \cdots, N \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \to w^T \varphi(x_i) + b + e_i - y_i = 0, i = 1, \cdots, N \end{cases} \tag{4}$$

After eliminating variables of (w, e), we obtain the solution

$$\begin{bmatrix} 0 & 1_v^T \\ 1_v & \Omega + \dfrac{1}{\gamma} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ y \end{bmatrix} \tag{5}$$

where

$y = [y_1, \cdots, y_N]^T; 1_v = [1, \cdots, 1]^T; \alpha = [\alpha, \cdots \alpha_N]$ and $\Omega_{il} = \varphi(x_i)^T \varphi(x_l)$ for $i =, l = 1, \cdots, N$. According to Mercer's condition, there is a mapping φ and kernel function $K(x_i, x_l) = \varphi(x_i)^T \varphi(x_l)$ with the relationship of

$$y(x) = \sum_{i=1}^{N} \alpha_k K(x, x_i) + b \tag{6}$$

where α and b are the solution to Eq. (4).

The kernel function $K(x, x_i)$ is any symmetric function that satisfies Mercer's condition. The typical examples of kernel function include linear, polynomial, radial basis function (RBF) kernel.

III. Hyper-parameter selection based on ACO

A Ant colony optimization (ACO)

ACO belongs to the class of biologically inspired heuristics. The basic idea of ACO is to imitate the cooperative behavior of ant colonies. ACO for solving combinatorial optimization problems was put forward by Colorni et al. [19]. The principle of these methods is based on the way that ants search for food and find their way back to the nest. During trips of ants a chemical trail called pheromone is left on the ground. The role of pheromone is to guide the other ants towards the target point. For one ant, a path is chosen according to the quantity of pheromone [20].

In ACO [21], a finite size colony of artificial ants is created. Each ant builds a solution. While building its own solution, each ant collects information based on the problem characteristics and its own performance. The performance measure is based on a quality function $F(\cdot)$. The ACO method can be applied to discrete combinational problems, where solutions to the optimization problem can be expressed in terms of feasible paths on a graph. Among all feasible paths, ACO aims to locate the one with a minimum cost. The problem of selecting the consequent of fuzzy rules can be designed as a combinational problem and solved by ACO. The information collected by the ants during the search process is stored in pheromone trails s associated to the connection of all edges. The ants cooperate in finding the solution by exchanging information via the pheromone trials. Edges can also have an associated heuristic value g. The g value represents a priori information about the problem instance definition or run-time information provided by a source different from the ants. The heuristic information is auxiliary in ACO and ACO works even without the use of it. Once all ants have completed their tours (i.e., at the end of the each iteration), ACO algorithms update the pheromone trails using all the solutions produced by the ant colony.

The whole ACO algorithm can be described by taking

the traveling salesman problem (TSP) as an example. The TSP is to find a minimal length with each city visited once. We are given a set of

N cities, represented by nodes, and a set E of edges with fully connecting nodes N. Let l_{ij} be the length of the $edge(i, j)$, that is the distance between cities i and j, with $i, j \in E$. At each iteration t, an ant in city i has to choose the next city j to head for from among those cities that it has not yet visited. The probability of picking a certain city j is calculated using the distance between cities i and j, and the amount of pheromone on the edge between these two cities. Ant colony system (ACS) algorithm was introduced to improve the performances of the basic algorithm [22] on big size problems [23]:

B ACO-based hyper-parameter selection

There are two key factors to determine the optimized hyperparameters using ACO: one is how to represent the hyper-parameters as the particle's position, namely how to encode.

Another is how to define the fitness function which evaluates the goodness of an ant. These two key factors are given as follows:

Encoding hyper-parameters: the optimized hyper-parameters for LS-SVMs include kernel parameter and regularization parameter. In solving hyper-parameter selection by the ACO, each ant is requested to represent a potential solution, namely hyper-parameters combination. So let us denote an m-hyper-parameters combination as a vector of dimension m. For example, if radial basis function is chosen as a kernel function, we denote the vector as (γ, σ), where γ is the regularization parameter, σ are kernel parameter.

Fitness function: the fitness function is the generalization performance measure. There are some different descriptions for the generalization performance measure.

In ACO, the fitness value is used to evaluate goodness of the particles, namely hyper-parameter combination. An ideal fitness value should reflect the generalization performance of LS-SVMs for different hyper-parameter combination.

The k-fold cross-validation method is used to define the fitness value. K-fold cross validation is one way to improve over the holdout method. The data set is divided into k subsets, and the holdout method is repeated k times. Each time, one of the k subsets is used as the test set and the other k-1 subsets are put together to form a training set. Then the average error across all k trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set k-1 times. The variance of the resulting estimate is reduced as k is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch k times, which means it takes k times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set k different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over [24].

In the simulations, k is five. To define the fitness value, we perform the five-fold cross validations with data in the training set for each particle and the average correct rate is taken as the fitness value, denoted as $Avetest_{-5}$ [25].

Therefore, the corresponding fitness can be determined. The fitness of a particle is evaluated by the following formulation:

$$f_i = Avetest_{-5} \tag{7}$$

where f_i is the fitness of ant i, and $Avetest_{-5}$ is the fitness value.

V. Results and discussion

A Selection of the optimal LS-SVMs parameter

Some 600 sequences of outer membrane proteins are annotated in the Swiss Prot database, 400 of which are from bacteria.

The precision and convergence of LS-SVM are affected by regularization parameter γ and kernel width σ. For the effective application of LS-SVMs, there is a need for a method to estimate these two parameters. So in order to obtain high level for predicting transmembrane β strands, γ and σ in the LS-SVMs have to be tuned by ACO method.

The optimal values for the regularization parameter γ and the kernel parameters σ with RBF kernel are shown in Fig. 1. In the final optimal LS-SVMs parameters are: γ = 1311.9, σ = 0.6655.

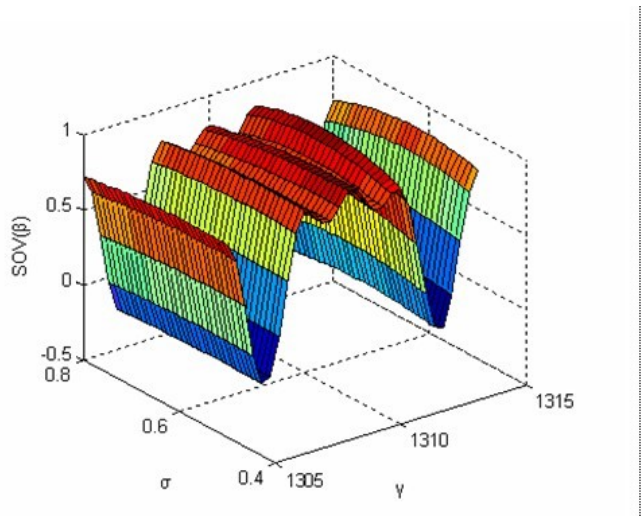

Fig. 1 The optimal values for the regularization parameter γ and the kernelparameters σ of LS-SVM with RBF kernel

B Simulation results of the testing database

We have predicted the transmembrane β strands for the testing database and compared our results with SVM and HMM method. The results are presented in table 1. From this table, we observe that the present method for LS-SVMs with RBF kernel predicts the transmembrane β

strands with higher accuracy than SVM and HMM method. These indicate LS-SVM is a powerful tool for predicting transmembrane β strands and our LS-SVMs model presented in this paper is accurate and valid.

Table 1 Comparisons of the performance between ACO-based hyper-parameter selection for LS-SVMs and other methods.

	Q_2	$Q(\beta)$	$Q(c)$	$P(\beta)$	$P(c)$	$C(\beta)$	$Sov(\beta)$
LS-SVM Testing	0.90	0.86	0.88	0.87	0.86	0.83	0.89
SVM-Testing	0.82	0.83	0.82	0.81	0.78	0.73	0.81
HMM testing	0.78	0.75	0.72	0.74	0.76	0.68	0.76

VI. Conclusion

We develop a least-squares support vector machines (LS-SVMs) model, which can predict the topology of β membrane proteins. LS-SVM has the problem of parameter selection. The selection of hyper-parameters plays an important role to the performance for transmembrane β strands prediction of LS-SVMs. The ACO method was proposed to select hyper parameter of LS-SVM. Simulation results are provided for showing the efficiency of the proposed method. From experimental result we observe that the present method for LS-SVMs with RBF kernel predicts the transmembrane β strands with higher accuracy than SVM and HMM method. Research results show that this model has high prediction precision, and it satisfies the need of β membrane proteins prediction.

Acknowledgement

The authors acknowledge the support of the South China Normal University and South China University of Technology. The work was support by the project of research of support vector machine in classification and regression, under project number Guangdong financial education (2008) 342. The work was also support by Guangdong province natural science fund (project No. 8151063101000040).

REFERENCES

[1] Pier Luigi Martelli, Piero Fariselli, Anders Krogh, and Rita Casadio. A sequence-profile-based HMM for predicting and discriminating β barrel membrane proteins. Bioinformatics, vol. 18 suppl.1 2002 s46-s45.

[2] White, S.H. and Wimley, W.C. Membrane protein folding and stability:Physical principles. Annu. Rev. Biophys. Biomol. Struct. 1999 28: 319–365.

[3] Schulz, G.E. 2000. β-barrel membrane proteins. Curr. Opin. Struct. Biol. 10: 443–447.

[4] Cowan, S.W. and Rosenbusch, J.P. 1994. Folding pattern diversity of integral membrane proteins. Science 264: 914–916.

[5] Gouaux, E. 1998. Roll out the barrel Nat. Struct. Biol. 5: 931–932.

[6] Irene Jacoboni, Pier Luigi Martelli, Piero Fariselli, Vito De Pinto, and Rita Casasio. Prediction of the transmembrane regions of β-barrel membrane proteins with a neural network-based predictor. Protein Science (2001), 10:779–787.

[7] Navjyot K. Natt, Harpreet Kaur, and G. P. S. Raghava. Prediction of Transmembrane Regions of β-Barrel Proteins Using ANN- and SVM-Based Methods. PROTEINS: Structure, Function, and Bioinformatics 56:11–18 (2004).

[8] Zou L, Wang Z, Wang Y. Prediction of outer membrane proteins using support vector machine with combined features. Sheng Wu Gong Cheng Xue Bao. 2008 Apr;24(4):651-8. Chinese.

[9] V. Vapnik, The Nature of Statistical Learning Theory,Wiley, New York, USA, 1995.

[10] V. Vapnik, S. Golowwich, A. Smola, Support vector method for function approximation, regression estimation and signal processing, Adv. Neural Inform. Process. Systems MIT Press 9 (1997) 281–287.

[11] G.W. Flake, S. Lawrence, Efficient SVM Regression Training with SMO, Machine Learning, Vol. 46, Kluwer Academic Publishers, Netherlands, 2002, pp. 271–290.

[12] B. Scholkopf, A.J. Smola, R.C. Williamson, P.L. Bartlett, New support vector algorithms, Neural Comput. 12 (4) (2000) 1207–1245.

[13] Smola, A.J. and B. Schölkopf: A Tutorial on Support Vector Regression. Statistics and Computing 14(3), 199-222 (2004).

[14] Wen Wen, Zhifeng Hao, Xiaowei Yang. A heuristic weight-setting strategy and iteratively updating algorithm weighted least-squares support vector regression. Neurocomputing 71 (2008) 3096– 3103.

[15] J.A.K. Suykens, J. Vandewalle, Least squares support vector machine classifiers, Neural Process. Lett. 9 (3) (1999) 293–300.

[16] J.A.K. Suykens, L. Lukas, J. Vandewalle, Sparse approximation using least squares support vector machines, in: Proceedings of the IEEE International Symposium on Circuits and Systems (ISCAS 2000), vol. 2, 2000, pp. 757–760.

[17] X.C. Guo, J.H. Yang, C.G. Wu, C.Y. Wang, Y.C. Liang. A novel LS-SVMs hyper-parameter selection based on particle swarm optimization. Neurocomputing 71 (2008) 3211– 3215.

[18] Jianping Wang, Eseosa Osagie, Parimala Thulasiraman, Ruppa K. Thulasiram. HOPNET: A hybrid ant colony optimization routing algorithm for mobile ad hoc network. Ad Hoc Networks 7 (2009) 690–705.

[19] Colorni A, Dorigo M, Maniezzo V. The ant system: an autocatalytic process. Technical report no. 91-016, Politecnico di Milano, Italy, 1991.

[20] Saadettin Erhan Kesena, M. Duran Toksari, Z¨ ulal G¨ ung¨ orc,ErtanG¨ unerc. Analyzing the behaviors of virtual cells (VCs) and traditionalmanufacturing systems: Ant colony optimization (ACO)-basedmetamodels. Computers & Operations Research 36 (2009) 2275 – 2285.

[21] Dorigo, M., & Stützle, T. (2004). Ant colony optimization. MIT. July.

[22] Dorigo M, St ¨ utzle T. Ant colony optimization. Cambridge, MA: MIT Press; 2004.

[23] Dréo J, Pétrowski A, Siarry P, Taillard E. Métaheuristiques pour l'optimisation difficile. Eyrolles; 2003.

[24] http://en.wikipedia.org/wiki/Cross-validation.

[25] P. Meinicke, T. Twellmann, H. Ritter, Discriminative densities from maximum contrast estimation, in: S. Becker, S. Thrun, K. Obermayer (Eds.), Advances in Neural Information Processing Systems, vol. 15, MIT Press, Cambridge, 2002, pp. 985–992.