COMPARISON OF K-NN AND SVM IN CLASSIFICATION OF PROTEINS INDUCED DURING SOMATIC EMBRYOGENESIS IN COCOA

G. L. Sairam, M. K. Rajesh and George V. Thomas

Bioinformatics Centre, Central Plantation Crops Research Institute, Kasaragod, Kerala

Introduction

In genomic research, classification of proteins into existing categories is used to learn the functions of new proteins. There are three major challenges in sequence classification. Firstly, most of the classifiers, such as decision trees and neural networks, can only take input data as a vector of features. However, there are no explicit features in sequence data. Secondly, even with various feature selection methods if we can transform a sequence into a set of features, the feature selection is far from trivial. The dimensionality of the feature space for the sequence data can be very high and the computation can be costly. Lastly, besides accurate classification results, some applications may require an interpretable classifier. Building an interpretable sequence classifier is difficult since there are no explicit features.

Recently, approaches based on distance-based classifiers have been used to predict the sub-cellular localization of proteins (Aarti *et al.,* 2005), recognize protein folds and super family (Iain *et al.,* 2007; Yang *et al.,* 2007), predict the alpha turn types, protein secondary structure prediction (Ashish and Bijnan, 2008) and to predict membrane protein types (Hongbin and Kuochen, 2005). In this work, we have developed a method for classifying proteins induced during somatic embryogenesis in cocoa based on amino acid and dipeptide composition. For both classifications task, the machine learning software Weka (Mark *et al.,* 2009) was used and for LibSVM, the separate jar file was collected and included in Weka.

Materials and Methods

Dataset

Cocoa SERK and LEC sequences were obtained from NCBI (National Center for Biotechnology Information). To reduce the redundancy in the data, Expasy sequence alignment tool- Decrease Redundancy (http://expasy.org/tools/redundancy/)- was used with a criteria that no two sequence had >90% sequence identity to any other sequence in the data set. The final data set contained 95 non-redundant set of these proteins from plants. The 95 sequences from plants were composed of 51 of Somatic Embryogenesis Receptor-like Kinase (SERK) and 44 of Leafy Cotyledon (LEC).

Support vector machine

Support Vector Machine (SVM) is a set of related supervised learning methods used for classification and regression. Viewing input data as two sets of vectors on an n-dimensional space, an SVM will construct a separating hyperplane in that space, one which maximizes the margin between the two data sets. To calculate the margin, two parallel hyperplane are constructed, one on each side of the separating hyperplane, which are "pushed up against" the two data sets (Vapnik, 1995). SVM modules were implemented by using LibSVM, the library of support vector machines (SVM) and the main goal is that users can use SVM as a tool. The different formulations included in LibSVM are C-support vector classification (C-SVC), v-support vector classification (v-SVC), distribution estimation (one-class VM), e-support vector regression (e-SVR), and v-support vector regression (v-SVR) (Chang and Lin, 2001).

KNN (k - nearest neighbour)

K-nearest neighbor (KNN) is a supervised learning algorithm where the output of test data is classified based on the majority of k-nearest neighbour category. The classification is mainly based on the Euclidean distance between a test sample and the training sample. The classification is mainly based on the majority vote among the classification of the K objects (Yang *et al.,* 2007).

Amino acid composition

Amino Acid Composition (AAC) is the amount of each amino acid present in the protein sequence, which transforms the protein sequence into an input vector of 20 dimensions. If Qi is the frequency of occurrence of an amino acid i, then the amino acid composition is AACi = Qi/L, where i is any amino acid.

$$\text{Fraction of aa (i)} = \frac{\text{Total number of amino acid of type (i)}}{\text{Total number of amino acids in protein}}$$

Equation 1: Amino Acid Composition Calculation

Dipeptide composition

For Dipeptide Composition (DC), the protein sequence is represented into an input vector of 400 (20 * 20) dimensions. L is the total number of all possible dipeptide in protein P (L = 400). Then Qij be a fraction of any pair of amino acids i, j is be any amino acid from 1 to 20. The dipeptide composition DCij= Qij/L

$$\text{Fraction of dipeptide (i)} = \frac{\text{Total number of dipeptide of type (ij)}}{\text{Total number of all possible dipeptide}}$$

Cross validation

A 5-fold LOOCV (Leave-One-Out Cross Validation) was carried out to evaluate the performance of SVM in the training set. In cross validation experiment the data set is divided into 5 equal sized samples and for each experiment use 5-1 samples for training and the remaining samples for testing. The accuracy was calculated as the average accuracy over 5 samples.

Validation test

The training set containing 95 sequences was used to evaluate the performance of SVM. A confusion matrix with True Positive, True Negative, False Positive and False Negative was created to determine the performance of SVM on training set. We also calculated sensitivity, specificity, accuracy and MCC (Mathew's correlation coefficients) from the performance results. Figure 1 shows the calculation of different parameters.

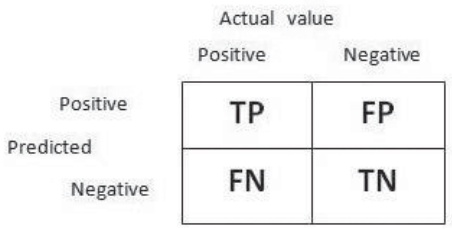

Fig. 1. Classification of a prediction into True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). TP known positive data predicted as positive; TN known negative data predicted as negative; FP known negative data predicted as positive; FN known positive data predicted as negative.

Results and Discussion

LibSVM and kNN modules are developed for classifying proteins induced during somatic embryogenesis in cocoa using amino acid composition and dipeptide composition. Both modules are trained using 5 fold Leave-one-out cross-validation. The LibSVM module developed using amino acid composition shows a classification accuracy of 90% with a Matthew's correlation (MCC) value of 0.81 with 100% specificity 83% sensitivity. The kNN module developed using amino acid composition shows a classification accuracy of 100% with a MCC value of 1.00 with 100% specificity and 100% sensitivity. The LibSVM module developed using dipeptide composition shows a classification accuracy of 100% with a MCC value of 1.00 with 100% specificity and 100% sensitivity. The kNN module developed using dipeptide composition shows a classification accuracy of 90% with a MCC value of 0.81 with 100% specificity and 83% sensitivity. Table 1 shows the detailed accuracy on classification.

Table 1. Detailed accuracy obtained from SERK and LEC proteins

Classifier / Prediction Method	LibSVM	kNN
Amino Acid composition	90%	100%
Dipeptide Composition	100%	90%
Average	95%	95%

The results showed that both distance based classifiers (LibSVM and kNN) have high classification accuracy on sequential data. For LibSVM the kernel used was RBF (Radial Basis Function) and the value fixed to 0.01. For kNN the k value was set to 1 and the distance function used was Euclidean distance. For all the classifier the cross validation used was LOOCV to get the classification accuracy. Figure 2 shows the average amino acid composition of SERK and LEC proteins.

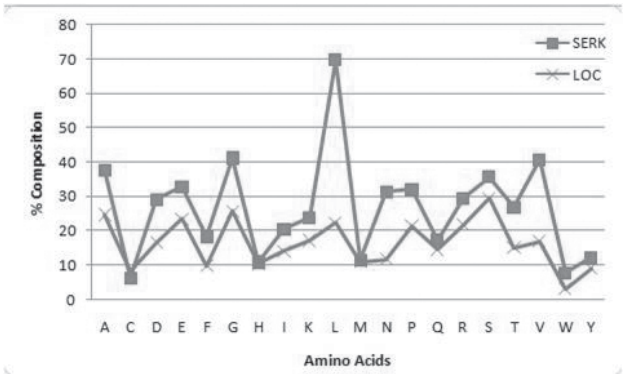

Fig. 2. Average amino acid composition of SERK and LEC proteins

Acknowledgement

This work was supported by a grant from Department of Biotechnology (BTISnet), New Delhi, India.

REFERENCES

1. Aarti, G., Manoj, B. and Raghava, G.P.S. 2005. SVM-based method for subcellular localization of human proteins using amino acid composition, their order and similarity search. *J.Biol.Chem.* 280: 14427-32.
2. Ashish, G. and Bijnan, P. 2008. Protein secondary structure prediction using distance based classifiers. *International Journal of Approximate Reasoning* 47(1): 37-44.
3. Chang, C.C. and Lin, C.J. 2001. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm.
4. Hongbin, S. and Kuo-Chen, C. 2005. Using optimized evidence-theoretic K-nearest neighbour classifier and pseudo-amino acid composition to predict membrane protein types. *Biochemical and Biophysical Research Communications* 334(1): 288-292.
5. Iain, M., Eugene, le., Rui, K., Jason, W., William, S.N. and Chrostina, L. 2007. SVM-Fold: a tool for discriminative multi class protein fold and superfamily recognition; *BMC Bioinformatics* 8(Suppl 4): S2. [PMID: 1892081].
6. Mark, H., Eibe, F., Geoffrey, H., Bernhard, P., Peter, R., and Ian, H.W. 2009. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* 11(1).
7. Vapnik, V.N. 1995. *The nature of statistical learning theory*. Springer, New York.
8. Yang, S., Jian, H., Ding, Z., Hongyuan, Z. and Lee, G.C. 2007. IKNN: Informative K-Nearest Neighbour Pattern Classification. *Springer Berlin/ Heidelberg*. 4702: 248-264.