Lab Note

Alternative splice site recognition based on a new fuzzy support vector machine

Xiaoxia Li*,†, Bo Sun†, and Chengguo Li

School of Information Engineering, Southwest University of Science and Technology, Mianyang, Sichuan 621010, China
*†These authors contributed equally to this work.
*Correspondence address. Tel/Fax: +86-816-6088219; E-mail: lixiaoxia@swust.edu.cn

Accurate alternative splice site (ASS) recognition is an important and difficult topic in the gene identification, and the average recognition rate is still <85% [1]. Many statistical pattern recognition methods, such as neural networks (NNs) and support vector machine (SVM), were used for this task [2,3]. Among them, SVM can construct a good high-dimensional learning model in the case of limited training set size and has good generalization ability, which exhibits many unique advantages in solving the small sample, non-linear, and high-dimensional pattern recognition problems [4]. To reduce the impact of noise samples on constructing optimal hyperplane, fuzzy SVM (FSVM) method was proposed [5]. Each sample was assigned a different membership and had different contributions to the objective function. Because the noise samples had smaller memberships, and their effects on the separating hyperplane were reduced or eliminated.

The fuzzy membership function (FMF) design is critical for FSVM [6,7]. A good FMF should be able to assign support vectors higher membership while noise samples lower membership. FMF is generally constructed by the distances between samples and class centers [8], tightness defined by mixed kernel function [9], or tightness defined by mix kernel function in feature space [10]. These methods can reduce to some extent the impact of the noise samples, but also reduce the memberships of support vectors. Here we designed a new membership calculation method that can simultaneously reduce the noise sample memberships and increase the support vector memberships.

Given a fuzzy training samples set: 
\[ S = \{(x_1, y_1, s_1), (x_2, y_2, s_2), \ldots, (x_n, y_n, s_n)\}, \]
where \( x_i \) is sample vector, \( y_i \in \{-1, 1\} \) is the sample category label, \( 0 \leq s_i \leq 1 \) is the sample fuzzy membership that reflects the importance of \( x_i \). The FSVM decision function is

\[ f(x) = \text{sgn}\left\{ \sum_{i=1}^{m} a_i^* y_i K(x_i, x) + b^* \right\} \tag{1} \]

where \( K(x_i, x) = (x_i^T x + 1)^2 \) is the kernel function, \( a_i^* > 0 \) is the Lagrange multiplier corresponding to the support vector, and \( b^* \) is a threshold.

Here, distance-based FMF has been improved by taking into account the relative positions of a sample between two cluster centers and the tightness of space around the sample. Based on the assumption that support vector is closer to the decision interface, the initial membership was defined as distances ratio between the sample and the two cluster centers of positive and negative training sample sets. The sample tightness was calculated by K-nearest neighbor (KNN) method, which can reduce the membership of the noise samples. The ultimate membership was the multiplication of the initial membership and the tightness.

Assuming that \( S_+ \) and \( S_- \) were the positive and negative training sample sets, respectively, with \( n_+ \) and \( n_- \) samples, \( O_+ \) and \( O_- \) were positive and negative training sample centers, \( d_i^+ \) and \( d_i^- \) were distances between a positive sample to positive and negative samples centers, respectively:

\[ d_i^+ = ||x_i - O_+||, \quad x_i \in S_+ \]
\[ d_i^- = ||x_i - O_-||, \quad x_i \in S_- \]

\( d_i^+ \) and \( d_i^- \) were defined in the similar way:

\[ d_i^+ = ||x_i - O_+||, \quad x_i \in S_- \]
\[ d_i^- = ||x_i - O_-||, \quad x_i \in S_- \]

The initial fuzzy membership was constructed by

\[ s_i^+ = \frac{d_i^+}{d_i^-}, \quad s_i^- = \frac{d_i^-}{d_i^+} \tag{4} \]

This membership construction method makes a support vector obtain higher membership and has a greater contribution to optimal separating hyperplane. However, this approach also assigned the noise samples close to the separating hyperplane a larger membership.

KNN method was used to reduce the noise sample memberships. The distances between a positive sample and other
samples were defined as $d_{ij}^+$, and placed in ascending order:
\[
d_{ij}^+ = ||x_i - x_j||, \quad x_i, x_j \in S_+, \quad i \neq j
\]
\[
d_{i1}^+ \leq d_{i2}^+ \leq \cdots \leq d_{in}^+
\]
Mean distance of the first $k$ neighbors was calculated by
\[
b_i = 1/\left(\frac{1}{k} \sum_{j=1}^{k} d_{ij}^+\right)
\]
Tightness of sample $x_i$ was defined as
\[
u_i^+ = b_i/B
\]
where $B = \max(b_1, b_2, \ldots, b_{n_+})$. The tightness of negative samples $u_i^-$ can be calculated in the same way, both maximum values are 1. Finally, a sample membership is defined as follows:
\[
s_i = \begin{cases} 
    s_i^+ u_i^+, & x_i \in S_+ \\
    s_i^- u_i^-, & x_i \in S_-
\end{cases}
\]
Since the distances between the noise samples and the class centers are larger than other normal samples, they may receive higher initial memberships. But their tightness is less than the normal samples, the final membership will be reduced by multiplying with the tightness, and the support vectors will not be affected and maintain a higher membership.

Figure 1 shows the membership distribution of neighbor FSVM (NFSVM) on an artificial sample set randomly generated by normal distribution. The positive and negative sample normal distribution mean values are $(2, 2)$ and $(3.5, 3.5)$, the variances are both 1. There are 20 samples in each sample set.

In Fig. 1, the memberships of most samples in the cross-region are higher than the sample memberships in other regions. Positive sample memberships in cross-region are 1.2, 0.9, 0.8, 0.5, 0.4, 0.2, etc., which are far greater than the positive sample membership average of 0.33. Negative sample memberships in cross-region are 0.8, 0.5, 0.2, etc., also greater than the negative sample membership average of 0.2. Since cross-regional samples are potential support vectors, and will be used to construct an optimal separating hyperplane. It suggests that the proposed membership design method can assign support vector samples higher membership, improving the capability to distinguish support vectors and noise samples.

Human ASS data were extracted from ASD (Alternative Splicing Database) (http://www.ebi.ac.uk/asdaltsplice/) [11], which is based on the Ensembl 36.35i gene data for humans. Alternative splice events occurred at 3’ and 5’ were annotated with II-3P and II-5P, respectively. Standard overlapping introns were excluded. Positive sample sets were generated from −5 to +5 splicing sites for GT donor sites and −20 to +5 for AG receptor sites. Negative sample sets were generated from splice site flanking sequences. Negative samples were selected from the nearest GT/AG non-splice site, which is located 200 bp outside the ASS (GT upstream, AG downstream). The first 1/20 of total set were selected as experiment samples, including 4596 AG and 4924 GT sites. The 1/2 randomly selected samples were used as a training set, the left as an independent test set. Five-fold cross-validation method was used in SVM training process.

Figure 2 shows the NFSVM ASS prediction receiver operating characteristic curve (ROC), which was compared with distance FSVM (DFSVM) and feature FSVM (FFSVM).

In Fig. 2, the NFSVM proposed here had larger area under the curve than other FSVMs for human ASS recognition. The total accuracy achieved was 93.5% and 83.6% for GT donor and AG receptor splice site recognition, respectively.

In summary, unreasonable membership function design of FSVM will result in poor prediction accuracy because the memberships of support vectors are decreased. This report presented a new fuzzy membership design method taking into account the relative positions of samples between two class centers and the tightness of the space around the sample, which distinguishes the support vectors and noise samples obviously. By assigning the potential support vectors greater memberships, the NFSVM not only preserve the merits of good generalization ability, high precision, and good balance, but also reduce the noise impact on the optimal separating hyperplane. In human ASSs identification study, the total recognition accuracy of the GT donor sites and AG acceptor sites achieved 93.5 and 83.6%, respectively; the overall performance of the NFSVM using this membership design method exceeded other two FSVM methods.
Funding

This work was supported by a grant from the Scientific Research Fund of Sichuan Provincial Education Department (No. 11ZB106) and SWUST Doctor Fund (053109).

References


Figure 2 Human ASS prediction ROC by FSVM (A) GT donor sites. (B) AG receptor sites. NFSVM, DFSVM, and FFSVM represent neighbor, distance, and feature FSVM.