International Journal of Pure and Applied Mathematics
Volume 76 No. 3 2012, 439-447
ISSN: 1311-8080 (printed version)
url: http://www.ijpam.eu

ijpam.eu

FUZZY SUPPORT VECTOR MACHINES BASED ALGORITHM FOR PEPTIDE IDENTIFICATION FROM TANDEM MASS SPECTRA

Ling Jian[1] §, Zunquan Xia[2]
[1,2]School of Mathematical Sciences
Dalian University of Technology
Dalian, 116024, P.R. CHINA
[1]College of Science
University of Petroleum
Qingdao, 266555, P.R. CHINA

Abstract: Shotgun tandem mass spectrometry-based peptide sequencing using programs such as SEQUEST allows high-throughput identification of peptides, which in turn allows identification of corresponding proteins. This paper present a novel machine learning method based on Fuzzy Support Vector Machines (Fuzzy SVMs) to discriminate between correct and incorrect identified peptides using SEQUEST search results. Through incorporating fuzzy membership this method can reduce the effect of noise in data points. Experiments show that this approach outperforms the traditional SVMs based technique, and it's an promising algorithm for peptide identification task.

AMS Subject Classification: 92B05
Key Words: peptide identification, SEQUEST, fuzzy support vector machines, mass spectrometry

1. Introduction

Peptide identification by tandem mass spectrometry (MS/MS) is widely used

Received: February 20, 2012

§Correspondence author

for high-throughput identification of proteins in complex biological samples [1]. In this MS/MS setting, peptides are associated with tandem mass spectra by search engines such as SEQUEST, Mascot, X! Tandem, etc, compare peptide sequences with each spectrum and rate the quality of the match (PSM) with a score. However, this approach often falsely identifies the peptides. Moore et al [2] suggested to employ a target-decoy protein database to identify the false discovery rate (FDR) and evaluate the accuracy of database search results. Decoy databases contain either reversed or shuffled protein sequences derived from the target protein database. The database search engine assigns an observed spectrum to either a target or a decoy peptide. The assignment of a decoy PSMs is considered incorrect. Nonetheless, the target PSMs are frequently incorrect and need validate further.

To improve the reliability of peptide identification, recently a variety of techniques including statistical significance re-estimation, e.g.,PeptideProphet [3], machine learning, e.g. Percolator [4] and [5], have been applied to post process the outputs of search engines. Generally speaking, based on the outputs of search engines these tools attempt to discriminate correct PSMs from incorrect PSMs. Thereinto Noble et al. [4] and [5] introduced SVMs into the peptide identification field and developed a SVMs-based method, called Percolator, for classifying PSMs data. SVMs proposed by Vapnik et al. [6] and [7] are widely used in real applications for promising classification performances. However, in setting of peptide identification problem decoy PSMs are labeled as negative points while target PSMs are labeled as positive points, as stated above the target PSMs are frequently incorrect so the points with positive label are not credible. So the peptide identification problem is not a classical binary classification problem. More attention should be payed to distinguish the effects of positive and negative points.

To combat this problem, we take the incorrect target PSMs as noise, and utilize the distances of positive points to the center of negative points to construct the fuzzy membership as one estimation of sample' confidence level. Fuzzy membership are further imported into Fuzzy SVMs proposed by Lin and Wang [8] to reduce the effects of noise. To heel, we apply Fuzzy SVMs based method and SVMs based method to refine the searching results of SEQUEST on a data set derived from synthetic protein mixtures. Performances comparison on various criteria show that the proposed Fuzzy SVMs method is a promising approach for peptide identification task.

The rest of this paper is organized as follows: Section 2 introduces briefly the algorithm of Fuzzy SVMs. Section 3 provides the application and validation of models to a PSMs data set, and the detailed analysis about the experimental

results is also given in this section. Finally, Section 4 concludes this paper.

2. Fuzzy Support Vector Machines

SVMs is a powerful tool for solving classical classification problems through constructing a hyperplane that can separate two classes by maximizing the margin [6] and [7]. It is clearly that all training points of each class are treated uniformly by SVMs. But in many application problems, such as peptide identification, the confidence level of training points are different, so the effects of different training points should vary one another. SVMs become invalid for every training points are treated comparably. To overcome this limitation of SVMs, Lin and Wang proposed Fuzzy SVMs to incorporate fuzzy membership associated with each training point into the SVMs [8].

Given a set of labeled training points with associated fuzzy membership $\mathbb{D} = \{(\mathbf{x}_i, y_i, s_i)\}_{i=1}^{l}$ with the vectors $\mathbf{x}_i$ from the pattern space $\mathbb{R}^n$ the classification label $y_i \in \{-1, 1\}$ and the fuzzy membership $0 \leq s_i \leq 1$ with $i = 1, \cdots, l$, the corresponding QP problem may be written as

$$\min_{\mathbf{w},b,\xi} \quad \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{l} s_i\xi_i \tag{2.1}$$

subject to

$$\begin{aligned} y_i(\mathbf{w}^{\mathrm{T}}\Phi(\mathbf{x}_i) + b) &\geq 1 - \xi_i, i = 1, \cdots, l \\ \xi_i &\geq 0, i = 1, \cdots, l \end{aligned} \tag{2.2}$$

where $\Phi(\mathbf{x})$ denotes the feature space vector, $\xi_i, i = 1, \cdots, l$ are slack variables measuring the degree of misclassification of the $\mathbf{x}_i$, and C is a positive constant controlling a trade off between a large margin and a small error penalty. Usually, for convenience of solving the above QP problem, Eqs. (2.1) and (2.2) are rewritten as the corresponding dual form by means of non-negative Lagrange multipliers $\alpha_i, i = 1, \cdots, l$, i.e.,

$$\begin{aligned} \max_{\alpha} \quad & \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j \Phi(\mathbf{x}_i)^{\mathrm{T}}\Phi(\mathbf{x}_j) \\ = \quad & \sum_{i=1}^{l}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \end{aligned} \tag{2.3}$$

subject to

$$\sum_{i=1}^{l} \alpha_i y_i = 0,$$
$$0 \leq \alpha_i \leq s_i C, \quad i = 1, \cdots, l. \tag{2.4}$$

Here, it should be noted that the inner product $\Phi(\mathbf{x}_i)^{\mathrm{T}}\Phi(\mathbf{x}_j)$ in Eq. (2.3) is denoted by $k(\mathbf{x}_i, \mathbf{x}_j)$, which is the well-known kernel function. It is usually specified beforehand as a function characterized by continuity, symmetry, and positive semidefiniteness, e.g., Gaussian radial basis function, sigmoid function, polynomial function, etc. The kernel trick provides a possibility of calculating the inner product in a high- or even infinite-dimensional feature space using low-dimensional pattern space data without knowing the exact form of Φ. Finally, solving Eqs. (2.3) and (2.4) can yield the optimal $\alpha_i, i = 1, \cdots, l$, and further yield the binary classifier as a function of the low dimensional pattern space data $\mathbf{x}$

$$f(\mathbf{x}) = \sum_{i=1}^{l} \alpha_i y_i k(\mathbf{x}_i, \mathbf{x}) + b. \tag{2.5}$$

3. Fuzzy Membership Design

In the current work, we utilize the distances of positive points to the center of negative points to construct the fuzzy membership and further import into Fuzzy SVMs to reduce the effects of noise. For positive samples, a native assumption is that the more close to the centroid of negative samples the more likely to be noise. Given a set $\mathbb{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{l}$ consisting of l^+ positive points and l^- negative points. Denote the set of positive points and negative as $\mathbb{D}^+ = \{(\mathbf{x}_i^+, 1)\}_{i=1}^{l^+}$ and $\mathbb{D}^- = \{(\mathbf{x}_i^-, -1)\}_{i=1}^{l^-}$, respectively. For description convenience, we denote the mapping of $\mathbf{x}$ in feature space as $\mathbf{z}$, i.e., $\mathbf{z} = \Phi(\mathbf{x})$. The centroid of negative points in feature space is noted by $\mathbf{z}_C = \frac{1}{l_-}\sum_{i=1}^{l_-} \Phi(\mathbf{x}_i{}^-)$. For a given positive sample $\mathbf{x}_j{}^+$, the distance of corresponding mapping $\mathbf{z}_j{}^+$ and the centroid of negative samples $\mathbf{z}_C{}^-$ can be get as

$$d_j \quad = \quad \|\Phi(\mathbf{x}_j^+) - \frac{1}{l_-}\sum_{i=1}^{l_-} \Phi(\mathbf{x}_i^-)\| \tag{3.1}$$

$$= \langle \Phi(\mathbf{x}_j^+) - \frac{1}{l_-}\sum_{i=1}^{l_-}\Phi(\mathbf{x}_i^-), \Phi(\mathbf{x}_j^+) - \frac{1}{l_-}\sum_{i=1}^{l_-}\Phi(\mathbf{x}_i^-)\rangle^{\frac{1}{2}} \quad (3.2)$$

$$= (k(\mathbf{x}_j^+, \mathbf{x}_j^+) - \frac{2}{l_-}\sum_{i=1}^{l_-}k(\mathbf{x}_j^+, \mathbf{x}_i^-) + \frac{1}{l_-^2}\sum_{i=1}^{l_-}\sum_{k=1}^{l_-}k(\mathbf{x}_i^-, \mathbf{x}_k^-))^{\frac{1}{2}} \quad (3.3)$$

Setting the fuzzy membership s_i of negative as 1 for the label of the negative point are credible, and employ the distances of positive point to the centroid of negative points to get the fuzzy membership of positive points as following

$$s_i = \frac{d_i}{\max(d_i)}, \quad i = 1, \cdots, l^+. \quad (3.4)$$

4. Performance Evaluation

4.1. Data Preprocess

In this section, a PSMs data set with 1076 positive samples and 1091 negative samples is selected to evaluate and compare the performance of Fuzzy SVMs with SVMs. The inputs of PSMs data are represented as the vector of the attributes from SEQUEST scores: (1)Xcorr, cross correlation between calculated and observed spectra, (2)deltaCn, the normalized difference between the Xcorr scores for the best and the second best scoring peptides, (3)SpRank, measures how the top scoring peptide ranked with respect to other candidate peptides during the preliminary scoring step, (4)Ions, the fraction of matched b and y ions, (5)Mass, the observed mass $[M + H]^+$. The basic statistical analysis is listed in Table 1. Due to the fact that the variable with a large magnitude

	Xcorr	DeltaCn	Ions	SpRank	Mass
Maximun	6.275	0.749	0.1	499	4658.5
Minimum	0	0	1	299.287	0.076
Mean	1.097	0.092	35.946	1247.137	0.371

Table 1: Statistical property of inputs

will have a stronger effect on the model parameters than the one with a small magnitude, it is ill-considered to take the data of these variables as the model

inputs for training. Thus, the following relationship is used to handle the input data to make them under the same magnitude.

$$u_i^j = \frac{x_i^j - m(x^j)}{\delta(x^j)}, \quad i = 1, \cdots, l; \quad j = 1, \cdots, k \tag{4.1}$$

where u_i^j, $m(x^j)$ and $\delta(x^j)$ stand for the actual jth model input, the mean and the standard deviation of the variable x^j, respectively.

In the following, the handled input data in Section 4.1 is fed into the Fuzzy SVMs and SVMs, respectively, to train the model parameters and further to perform peptide identification task.

4.2. Comparison of Fuzzy SVMs Results with SVMs

To evaluate the performance of the Fuzzy SVMs model and SVMs model quantitatively, the following frequently used criteria in the bioinformatic field are computed: true positive rate

$$TPR = TP/(TP + FN) \tag{4.2}$$

false positive rate

$$FPR = FP/(FP + TN) \tag{4.3}$$

false discovery rate

$$FDR = 2FP/(FP + TP) \tag{4.4}$$

where TP is the number of positive sample true classified and FP is the number of positive sample false classified, FN is the number of negative sample false classified and FT is the number of negative sample true classified. A confusion matrix is shown in Table 2 to facilitate instructions.

	Predicted class	
Actual class	Positive	Negative
Positive	TP	FN
Negative	FP	TN

Table 2: Confusion matrix of classification problem

The above evaluation criteria are used to evaluate the effectiveness of a proposed method. Table 3 presents the validated PSMs generated from each

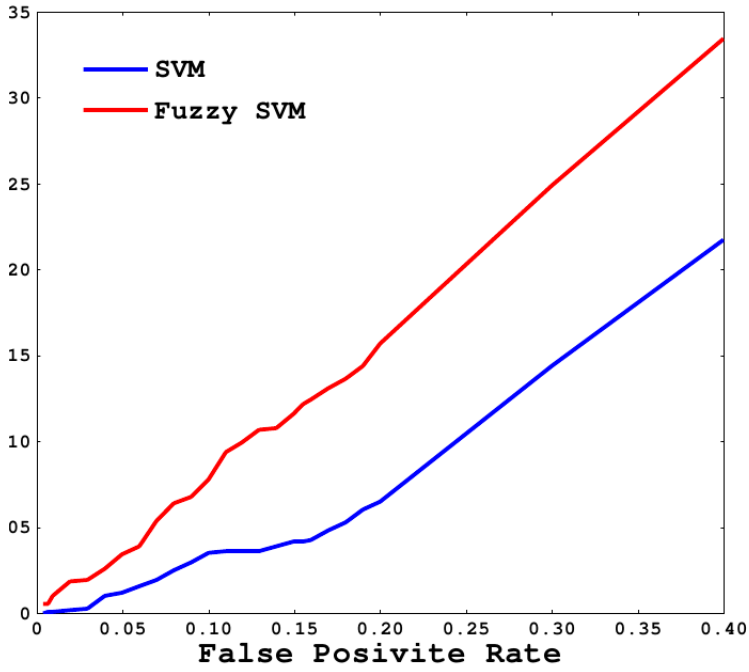

Figure 1: Performance Comparison of Fuzzy SVMs and SVMs

approach under the same settings of FPR. Lower FDR, higher TPR, and more TP all indicate that the Fuzzy SVMs based method is more suitable for the peptide identification task.

Receiver operating characteristic ROC plot is employed to further evaluate the performance of the two method. ROC plot is a graphical plot of the true positive rate vs. false positive rate for a binary classifier. Generally speaking, the closer the curve follows the left-hand border and then the top border of the ROC space, the better performance of the approach, which means points in ROC space are better than other points if they are plotted to the upper left corner.

ROC curves plotted in Figure 1 demonstrate that the compared with SVMs based method the Fuzzy SVMs based method improved sensitivity and specificity for distinguishing correct and incorrect PSMs significantly.

Algorithms	TPR	FDR	TP
Fuzzy SVMs	0.421	1.092	453
SVMs	0.316	1.232	340

Table 3: Comparison of Fuzzy SVMs and SVMs

5. Conclusion

In this paper, we propose a Fuzzy SVM based model to reduce the effect of noise for peptide identification task: firstly utilize the distances of target PSMs to the center of decoy points to construct the fuzzy membership, secondly impose the fuzzy membership into Fuzzy SVMs model such that different target PSMs points can make different contributions to the learning of decision surface. Comparison with SVMs based method on a synthetic protein mixture data set demonstrate the proposed Fuzzy SVMs based model is a highly sensitive and specific technique for peptide identifications.

References

[1] R. Aebersold, M. Mann, Mass spectrometry-based proteomics, *Nature*, **422** (2003), 198-207.

[2] R.E. Moore, M.K. Young, T.D. Lee, Qscore, An algorithm for evaluating sequest database search results, *Journal of the American Society for Mass Spectrometry*, **13** (2002), 378-386.

[3] A. Keller, A. I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, *Analytical Chemistry,* **74** (2002), 5383-5392.

[4] L. Kall, J.D. Canterbury, J. Weston, W.S. Noble, M.J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets, *Nature Methods*, **4** (2007), 923-925.

[5] A.A. Klammer, X. Yi, M.J. MacCoss, W.S. Noble, Improving tandem mass spectrum edentification using peptide retention time prediction across diverse chromatography conditions, *Analytical Chemistry*, **79** (2007), 6111-6118.

[6] N. Cristianini, J. Shawe-Taylor, *An Introduction to Support Vector Machines*, Cambridge, U.K.:Cambridge Univ. Press (2000).

[7] B. Schölkopf, A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*, Cambridge, MA: MIT Press (2002).

[8] C. Lin, S. Wang, Fuzzy support vector machines, *IEEE Transactions on Nerual Networks*, **13** (2002), 464-471.