Estimating the Class Posterior Probabilities in Biological Sequence Segmentation

Rémi Bonidal, Fabienne Thomarat, and Yann Guermeur

LORIA – Equipe ABC
Campus Scientifique, BP 239
54506 Vandœuvre-lès-Nancy Cedex, France
(e-mail: {Remi.Bonidal,Fabienne.Thomarat,Yann.Guermeur}@loria.fr)

Abstract. To tackle segmentation problems on biological sequences, we advocate the use of a hybrid architecture combining discriminant and generative models in the framework of a hierarchical approach. Multi-class support vector machines and neural networks provide a set of initial predictions. These predictions are post-processed by classifiers estimating the class posterior probabilities. The outputs of this cascade of classifiers, named MSVMpred, are then used to derive the emission probabilities of a hidden Markov model performing the final prediction. This article deals with the evaluation of MSVMpred both as a stand alone classifier, i.e., according to the recognition rate, and as part of a hybrid architecture, i.e., with respect to the quality of the probability estimates.
Keywords: Multi-class support vector machines, Class posterior probability estimates, Bioinformatics.

1 Introduction

We are interested in problems of bioinformatics which can be specified as follows: a biological sequence must be split into consecutive segments belonging to different categories given a priori. This class contains problems of central importance in biology such as protein secondary structure prediction, alternative splicing prediction, or the search for the genes of non-coding RNAs. To tackle it, we advocate the use of a hybrid architecture combining discriminant and generative models in the framework of a hierarchical approach. It was first outlined in [7], and was later developed by the community (see for instance [15]). Discriminant models compute estimates of the class posterior probabilities based on local information extracted from the sequence (or a multiple alignment). These estimates are then post-processed by generative models, to produce the final prediction. The class of problems of interest further suggests a specific organization of the discriminant models: a first set of classifiers performs predictions based on the content of a window sliding on the sequence, and a second set combines and filters these predictions. The most successful instance of this "cascade" architecture, MSVMpred [11], is built as follows: the first level predictions are made by multi-class support vector machines (M-SVMs) [10] and neural networks (NNs) [17], the second level ones by "combiners" selected according to their capacity.

The present work aims at assessing the potential of MSVMpred through comparative studies. The assessment regards both the recognition rate and the quality of the probability estimates. Experiments are performed on synthetic data and in protein secondary structure prediction. The performance of reference is provided by the standard pairwise coupling procedure [12] applied to the (post-processed) outputs of bi-class support vector machines (SVMs) [5]. The results highlight the potential of MSVMpred and by way of consequence our hybrid architecture. The organization of the paper is as follows. Section 2 is a general introduction to MSVMpred. The basic experimental protocol is detailed in Section 3. Section 4 provides results obtained on synthetic data. Experimental results in protein secondary structure prediction are exposed in Section 5, and we draw conclusions in Section 6.

2 MSVMpred

Let $\mathcal{X}$ be a non empty set and $Q \in \mathbb{N} \setminus [\![0,2]\!]$. $\mathcal{X}$ and $[\![1,Q]\!]$ are respectively the description space and the set of categories of a discrimination problem characterized by a $\mathcal{X} \times [\![1,Q]\!]$-valued random pair (X,Y) whose distribution is unknown. All the knowledge regarding this distribution is provided by a set of labelled data. This set is used to select in a given class of functions a function assigning a category to the descriptions with minimal probability of error (risk). When the learning problem is formulated in that way, MSVMpred is simply defined as a two-layer cascade of classifiers with domain $\mathcal{X}$ and range the unit $(Q-1)$-simplex. M-SVMs and NNs produce initial predictions which, after a possible post-processing, are exploited by classifiers of appropriate capacity estimating the class posterior probabilities.

2.1 Multi-class support vector machines

M-SVMs are multi-class extensions of the (bi-class) SVM which do not rely on a decomposition method. As models of pattern recognition, they are characterized by the specification of a class of functions and a learning problem.

Definition 1 (Class of functions $\mathcal{H}_{\kappa,Q}$). Let κ be a real-valued positive type function [3] on $\mathcal{X}^2$ and let $\left(\mathbf{H}_\kappa, \langle\cdot,\cdot\rangle_{\mathbf{H}_\kappa}\right)$ be the corresponding reproducing kernel Hilbert space. The class of vector-valued functions on which a Q-category M-SVM with kernel κ is based is the class

$$\mathcal{H}_{\kappa,Q} = \left(\mathbf{H}_\kappa \oplus \{1\}\right)^Q$$

where $\{1\}$ is the space of real-valued constant functions on $\mathcal{X}$.

$\mathcal{H}_{\kappa,Q}$ is a class of mulivariate affine functions on $\mathbf{H}_\kappa$. Indeed,

$$\forall h \in \mathcal{H}_{\kappa,Q},\ \forall x \in \mathcal{X},\ \ h(x) = \bar{h}(x) + b = \left(\left\langle \bar{h}_k, \kappa(x,\cdot)\right\rangle_{\mathbf{H}_\kappa} + b_k\right)_{1 \leqslant k \leqslant Q},$$

where $\bar{h} = \left(\bar{h}_k\right)_{1 \leqslant k \leqslant Q} \in \mathbf{H}_\kappa^Q$ and $b = (b_k)_{1 \leqslant k \leqslant Q} \in \mathbb{R}^Q$.

Definition 2 (Generic model of M-SVM, Definition 4 in [10]). Let $\mathcal{X}$ be a non empty set and $Q \in \mathbb{N} \setminus [\![0,2]\!]$. Let κ be a real-valued positive type function on $\mathcal{X}^2$ and let $\mathcal{H}_{\kappa,Q}$ be the class of functions induced by κ according to Definition 1. For $m \in \mathbb{N}^*$, let $d_m = ((x_i, y_i))_{1 \leqslant i \leqslant m} \in (\mathcal{X} \times [\![1,Q]\!])^m$ and $\xi \in \mathbb{R}^{Qm}$ with $\left(\xi_{(i-1)Q+y_i}\right)_{1 \leqslant i \leqslant m} = 0_m$. A *Q-category M-SVM* with kernel κ and training set d_m is a discriminant model trained by solving a convex quadratic programming problem of the form

Problem 1 (Learning problem of an M-SVM, primal formulation).

$$\min_{h,\xi} \left\{ \|M\xi\|_p^p + \lambda \sum_{k=1}^{Q} \left\|\bar{h}_k\right\|_{\mathbf{H}_\kappa}^2 \right\}$$

$$s.t. \begin{cases} \forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ K_1 h_{y_i}(x_i) - h_k(x_i) \geqslant K_2 - \xi_{(i-1)Q+k} \\ \forall i \in [\![1,m]\!],\ \forall (k,l) \in \left([\![1,Q]\!] \setminus \{y_i\}\right)^2,\ \ K_3 \left(\xi_{(i-1)Q+k} - \xi_{(i-1)Q+l}\right) = 0 \\ \forall i \in [\![1,m]\!],\ \forall k \in [\![1,Q]\!] \setminus \{y_i\},\ \ (2-p)\xi_{(i-1)Q+k} \geqslant 0 \\ (1-K_1)\sum_{k=1}^{Q} h_k = 0 \end{cases}$$

where $\lambda \in \mathbb{R}_+^*$, $(K_1, K_3) \in \{0,1\}^2$, and $K_2 \in \mathbb{R}_+^*$. M is a $Qm \times Qm$ matrix of rank $(Q-1)\,m$ such that for all i in $[\![1,m]\!]$, its column of index $(i-1)Q + y_i$ is equal to 0_{Qm}. $p \in \{1,2\}$ and if $p = 1$, then M is a diagonal matrix.

If the problem of deriving class posterior probability estimates from the outputs of an SVM, in the bi-class case, or a set of SVMs, in the multi-class case, has been frequently addressed, this is not the case for the M-SVMs.

2.2 Second level discriminant models

Several options are available to perform the second level predictions. Both the polytomous logistic regression (PLR) model [13] and the linear ensemble methods (LEMs) [8] estimate the class posterior probabilities. Under mild hypotheses, this property is shared by many NNs, among which the most common one, the multi-layer perceptron (MLP) (see for instance [17]). All three models are assessed separately in our experiments. We have introduced them in order of increasing capacity [9]. Indeed, the PLR is a linear separator. An LEM combines Q-category classifiers taking their values in the unit $(Q-1)$-simplex. To combine classifiers that do not exhibit this property, such as the M-SVMs, the introduction of an intermediate step of post-processing is required, which can give birth to a nonlinear separator (in the space where the outputs of the base classifiers live). At last, an MLP using a softmax activation function for the output units and the cross-entropy (CE) loss (a sufficient condition for its outputs to be class posterior probability estimates) is an extension of the PLR obtained by adding a hidden layer. The boundaries it computes are nonlinear in its input space. The availability of classifiers of different capacities for the second level of the cascade is an important feature

of MSVMpred. It makes it possible to cope with one of the main limiting factors to the performance of modular architectures: overfitting.

3 Experimental protocol

When computing polytomies with SVMs, the standard way to derive class posterior probability estimates consists in combining the one-against-one decomposition scheme with pairwise coupling [12]. For each $x \in \mathcal{X}$, the $\binom{Q}{2}$ outputs of the SVMs are post-processed by a parameterized sigmoid [16] to provide estimates of the probabilities $\mathbb{P}(Y = k \mid Y \in \{k, l\}, X = x)$. These estimates are then used to derive estimates of the probabilities $\mathbb{P}(Y = k \mid X = x)$ by means of a maximum likelihood procedure.

The implementation of MSVMpred involves the four main models of M-SVMs as base classifiers, i.e., the models of Weston and Watkins (WW), Crammer and Singer (CS), Lee and co-authors (LLW), and the M-SVM^2 (see [10] for a survey). Let $I_{Qm}(d_m)$ and $M^{(2)}$ designate two instances of M whose general term $m_{(i-1)Q+k,(j-1)Q+l}$ is respectively $\delta_{i,j}\delta_{k,l}(1 - \delta_{y_i,k})$ and $(1 - \delta_{y_i,k})(1 - \delta_{y_j,l})\left(\delta_{k,l} + \frac{\sqrt{Q-1}}{Q-1}\right)\delta_{i,j}$, where δ is the Kronecker symbol. The formulations of the aforementioned M-SVMs as instances of the generic model correspond to the values of the hyperparameters reported in Table 1.

M-SVM	M	p	K_1	K_2	K_3
WW-M-SVM	$I_{Qm}(d_m)$	1	1	1	0
CS-M-SVM	$\frac{1}{Q-1}I_{Qm}(d_m)$	1	1	1	1
LLW-M-SVM	$I_{Qm}(d_m)$	1	0	$\frac{1}{Q-1}$	0
M-SVM^2	$M^{(2)}$	2	0	$\frac{1}{Q-1}$	0

Table 1. Specifications of the four M-SVMs used as base classifiers in MSVMpred

Unless otherwise specified, the SVMs and M-SVMs implemented use a spherical Gaussian kernel. To post-process the outputs of the M-SVMs prior to presenting them in input of an LEM, we resorted to the PLR. Besides the pairwise coupling, the set of models providing the performance of reference is made up of the PLR, the MLP, and an M-SVM (precisely the CS-M-SVM) post-processed by the PLR. The MLP, as a universal approximator, has the potential to approximate the probabilities, and thus the Bayes classifier, arbitrarily well. Its performance is limited by the efficiency of the back-propagation algorithm. For a classifier g from $\mathcal{X}$ into the unit $(Q-1)$-simplex, the empirical CE measured on $((x_i, y_i))_{1 \leqslant i \leqslant m}$ is $-\frac{1}{m}\sum_{i=1}^{m}\sum_{k=1}^{Q}\mathbb{P}(Y = k \mid X = x_i)\ln(g_k(x_i))$ when the probabilities are known, $-\frac{1}{m}\sum_{i=1}^{m}\ln(g_{y_i}(x_i))$ otherwise. This quantity is used to assess the accuracy of the probability estimates. The significance of the differences in recognition rate is measured by means of the two sample proportion test.

4 Experiments on synthetic data

We used an instance of the 10-category problem studied in [6]. In this problem, the categories are equiprobable and their probability density distributions are 20-dimensional Gaussians. The means of the Gaussians are randomly generated from a uniform distribution on $[0,1]^{20}$. The covariance matrices are also random. For each of them, the eigenvectors are generated from a uniform distribution on the unit sphere of $\mathbb{R}^{20}$ subject to orthogonality constraints. The square-roots of the eigenvalues are drawn from a uniform distribution on $[0.01, 1.01]$. Our Monte Carlo estimates of the risk and entropy of the Bayes classifier: $\mathbb{E}_{(X,Y)}\left[-\ln\left(\mathbb{P}\left(Y \mid X\right)\right)\right] = \mathbb{E}_X\left[-\sum_{k=1}^{Q}\mathbb{P}\left(Y=k \mid X\right)\ln\left(\mathbb{P}\left(Y=k \mid X\right)\right)\right]$ are respectively 0.50% and 0.0142. For these experiments, the NN used as base classifier, in parallel with the M-SVMs, is an MLP. To train MSVMpred, the training set is split into two subsets. Three quarters of the data are used for the base classifiers and their possible post-processing, the rest for the combiner. This decomposition is also the one used for the pairwise coupling. On the contrary, in the case when the CS-M-SVM alone is post-processed by the PLR, since no combiner is to be trained, the M-SVM and its post-processing are respectively trained on the first and the second subset. The test set is made up of 60000 examples. Figure 1 displays the prediction accuracy as a function of the training set size m, for all the classifiers but the PLR (alone), discarded for lack of performance. For m large enough (over 5000), the three variants of MSVM-

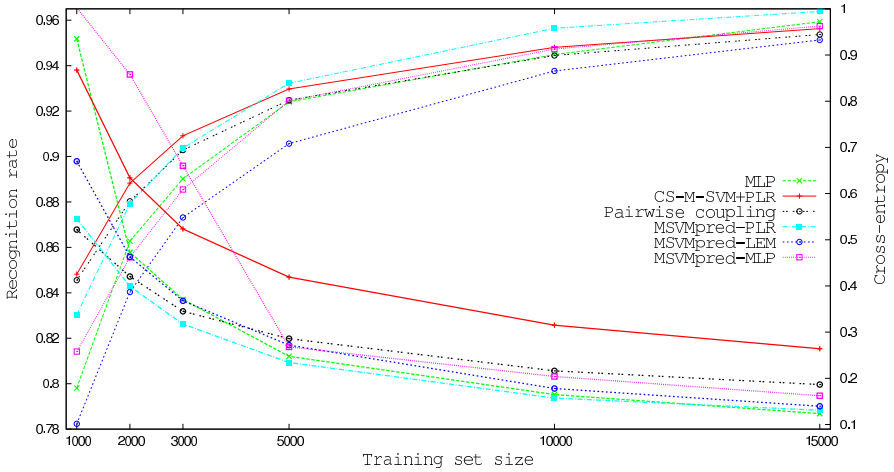

Fig. 1. Prediction accuracy as a function of the training set size

pred are uniformly superior to the pairwise coupling in terms of CE. The most efficient variant is the one of lowest capacity (PLR). For $m \geqslant 10000$,

its recognition rate is superior to that of the second best classifier with confidence exceeding 0.95. MSVMpred seems capable of making the best of the complementarity of the base classifiers, since this gain is not obtained at the expense of the quality of the probability estimates. Indeed, the CE of the best variant is inferior to that of the MLP (alone), until both values become indistinguishable (for $m \geqslant 15000$).

5 Protein secondary structure prediction

Predicting the secondary structure of a protein is a three-category discrimination task consisting in assigning a conformational state α-helix, β-strand or aperiodic (coil), to each residue (amino acid) of its sequence. In that context, the two levels of classification performed in cascade correspond respectively to a sequence to structure prediction and a structure to structure prediction. State-of-the-art prediction methods exhibit a recognition rate around 80%, whereas the best cascades published so far remain slightly below 78%.

5.1 Sequence-to-structure classification

For this classification, deriving the predictors from a multiple alignment rather than from the sole sequence of interest makes it possible to incorporate some evolutionary information. The descriptions of the residues of a protein sequence are thus derived from the corresponding position-specific scoring matrix (PSSM) produced by PSI-BLAST [1]. To generate the PSSMs, the version 2.2.25 of the BLAST package is used. Choosing BLAST in place of the more recent BLAST+ offers the facility to extract more precise PSSMs. Three iterations are performed against the nr database with an E-value inclusion threshold of 0.005. The nr database is filtered by pfilt [14] to remove low complexity regions, transmembrane spans and coiled coil regions. Since a sliding window is used, the description of a residue is thus obtained by appending consecutive rows of the PSSM associated with the sequence to which it belongs. To dedicate the M-SVMs to the task of interest, we chose an elliptic Gaussian kernel weighting differently the positions of this window. To set the values of the weights, we applied a straightforward multi-class extension of the kernel target alignment. We had to center the data in the feature space according to the formula given in [4], and to penalize the corresponding objective function with the ℓ_2 norm of the weighting vector. This kernel was also used by the SVMs involved in the pairwise coupling. To exploit the sequential nature of the data, the NN we used as additional base classifier was a recurrent one: the bidirectional recurrent neural network (BRNN) [2].

5.2 Structure-to-structure classification

As usual when applying structure-to-structure classification, the predictors are provided by the content of a window sliding on the outputs of the sequence-to-structure classifiers. For the PLR and the MLP specifically, they are

weighted according to their position in their window. The weighting is derived in the same way as the one of the elliptic Gaussian kernel.

5.3 Experimental results

The data set is CB513, a standard benchmark made up of 513 sequences for a total of 84119 residues. The secondary structure assignment was performed by the DSSP program, with the reduction from 8 to 3 conformational states following the CASP method. The first and second sliding window, centered on the residue of interest, have a size of 13 and 15 respectively. A seven-fold cross-validation procedure was implemented. At each step, two thirds of the training set were used to train the sequence-to-structure classifiers and their possible post-processing, and one third to train the combiner. This decomposition was also used for the pairwise coupling. Since a secondary structure prediction method must fulfill specific requirements in order to be useful for the biologist, two performance measures were added to the recognition rate (Q_3) and the CE: the Pearson correlation coefficients $C_{\alpha/\beta/\text{coil}}$ and the segment overlap measure (Sov'99). Results are gathered in Table 2.

	PLR	MLP	BRNN	CS-M-SVM + PLR	Pairwise coupling	MSVMpred			
						PLR	LEM	MLP	BRNN
Q_3 (%)	72.5	76.1	77.0	77.0	76.6	78.2	77.9	78.1	78.1
CE	0.674	0.585	0.568	0.578	0.581	0.542	0.551	0.537	0.548
C_α	0.62	0.71	0.72	0.72	0.71	0.74	0.74	0.74	0.74
C_β	0.55	0.61	0.62	0.62	0.61	0.64	0.64	0.64	0.64
C_{coil}	0.54	0.57	0.58	0.59	0.58	0.60	0.60	0.60	0.60
Sov'99 (%)	67.7	68.7	71.3	71.2	71.4	74.6	74.3	73.8	73.7

Table 2. Performance of MSVMpred in protein secondary structure prediction

The Q_3 of each variant of MSVMpred is statistically superior to that of the pairwise coupling and all the single classifiers with confidence exceeding 0.95. The value of the CE confirms this superiority. In comparison with the results obtained on the synthetic data, a nice feature is that all variants of MSVMpred perform equally well. Model selection is not an issue here.

6 Conclusions and ongoing research

This article has illustrated the potential of MSVMpred as discriminant model performing the low-level processing of the data in our hybrid architecture dedicated to biological sequence segmentation. In protein secondary structure prediction, its performance is slightly superior to that of the best cascades published so far, and significantly superior to that of pairwise coupling, or single classifiers. An appropriate choice of the combiner, governed by the concern of capacity control, could optimize jointly the recognition rate and the

quality of the class posterior probability estimates. We are currently working on increasing the scope of MSVMpred in biological sequence segmentation without loss of control on its generalization performance.

Acknowledgments The authors would like to thank C. Magnan for providing them with the 1D-BRNN package.

References

1. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs", *Nucleic Acids Research*, 25, 3389-3402 (1997).
2. Baldi, P., Brunak, S., Frasconi, P., Pollastri, G., and Soda, G., "Exploiting the past and the future in protein secondary structure prediction", *Bioinformatics*, 15, 937-946 (1999).
3. Berlinet, A., and Thomas-Agnan, C., *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, Boston (2004).
4. Cortes, C., Mohri, M., and Rostamizadeh, A., "Two-stage learning kernel algorithms", in *ICML'10*, (2010).
5. Cortes, C., and Vapnik, V., "Support-Vector networks", *Machine Learning*, 22, 273-297 (1995).
6. Friedman, J., "Another approach to polychotomous classification", Technical Report, Department of Statistics, Stanford University (1996).
7. Guermeur, Y., "Combining discriminant models with new multi-class SVMs", *Pattern Analysis and Applications*, 5, 168-179 (2002).
8. Guermeur, Y., "Ensemble methods of appropriate capacity for multi-class support vector machines", in *SMTDA'10*, 311-318 (2010).
9. Guermeur, Y., "Sample complexity of classifiers taking values in $\mathbb{R}^Q$, application to multi-class SVMs", *Communications in Statistics - Theory and Methods*, 39, 543-557 (2010).
10. Guermeur, Y., "A generic model of multi-class support vector machine", *International Journal of Intelligent Information and Database Systems*, (accepted).
11. Guermeur, Y., and Thomarat, F., "Estimating the class posterior probabilities in protein secondary structure prediction", in *PRIB'11*, 261-271 (2011).
12. Hastie, T., and Tibshirani, R., "Classification by pairwise coupling", *The Annals of Statistics*, 26, 451-471 (1998).
13. Hosmer, D.W., and Lemeshow, S., *Applied Logistic Regression*, Wiley, London (1989).
14. Jones, D.T., and Swindells, M.B., "Getting the most from PSI-BLAST", *Trends in Biochemical Sciences*, 27, 161-164 (2002).
15. Lin, K., Simossis, V.A., Taylor, W.R., and Heringa, J., "A simple and fast secondary structure prediction method using hidden neural networks", *Bioinformatics*, 21, 152-159 (2005).
16. Platt, J.C., "Probabilities for SV machines", in *Advances in Large Margin Classifiers*, Smola, A.J., Bartlett, P.L., Schölkopf, B., and Schuurmans, D., Eds., The MIT Press, Cambridge MA, 61-73 (2000).
17. Richard, M.D., and Lippmann, R.P., "Neural network classifiers estimate Bayesian *a posteriori* Probabilities", *Neural Computation*, 3, 461-483 (1991).