Spatially-constrained sample kernel for sequence classification

Pavel P. Kuksa, Pai-Hsi Huang, Vladimir Pavlovic

Department of Computer Science
Rutgers University
{pkuksa, paihuang, vladimir}@cs.rutgers.edu

Kernel-based learning methods provide some of the most accurate results in many sequence analysis and prediction tasks [1, 2, 4, 6]. However, the improved accuracy is often achieved at the cost of high computational complexity of training and prediction. We propose a new family of the string-based kernel classification methods for the sequence analysis tasks that offer low computational cost and display the state-of-the-art performance. We illustrate our approach on protein remote homology classification problems [2, 3, 5, 7] under supervised and semi-supervised settings.

In contrast to traditional string kernels, spatially-constrained sample kernels sample the sequence features at multiple resolutions, establishing the similarity measure across different scales, with potentially highly diverse mutation/insertion/deletion process. In particular, the kernels $K(\cdot,\cdot|k,t,d)$ have the following form

$$K(X,Y|k,t,d) = \sum_{\substack{a_1,\ldots,a_t \in \Sigma^k \\ d_1,\ldots,d_{t-1} \in \{0,1,\ldots,d-1\}}} C(a_1,d_1,a_2,d_2,...,d_{t-1},a_t|X)C(a_1,d_1,a_2,d_2,...,d_{t-1},a_t|Y)$$

where $C(a_1,d_1,a_2,d_2,...,d_{t-1},a_t|X)$ is the number of times substrings $a_1 \overset{d_1}{\leftrightarrow} a_2 \overset{d_2}{\leftrightarrow}, ..., \overset{d_{t-1}}{\longleftrightarrow} a_t$ (a_1 separated by d_1 characters from a_2 separated by d_2 characters from a_3 etc.) occurring in the sequence X. Each sample consists of t spatially-constrained probes a_i of size k, with the probes spatially constrained to lie no more than d positions away from their neighbors. In the proposed kernels, parameter k controls the individual probe size, parameter d controls the locality of the sample, and parameter t controls the cardinality of the sampling neighborhood. For instance, $k = 1, t = 2$ denote kernels constructed from pairs of monomers, $t = 3$ triples of monomers, etc.

The proposed sample string kernels, unlike the spectrum-like kernels (e.g. exact spectrum or mismatch kernels [4]), not only take into account the feature counts, but also include spatial configuration information, i.e. how the features are distributed in the sequence. The addition of the spatial information can be critical in establishing similarity of sequences under complex transformations such as the evolutionary processes in protein sequences. Interestingly, very short sequence features (we use only one-character-long monomers) achieve performance better than the standard string kernels with longer but spatially more constrained string features.

The possibility of using short features a_i can also lead to significantly lower computational complexity of computing the new kernels. The resulting feature spaces have substantially fewer dimensions: 2000 and 72000 for double and triple kernels, respectively, compared to 3200000 of the competing non-spatial string kernels (spectrum/mismatch kernels.) As a consequence, the proposed kernels can be efficiently computed using a sorting and counting approach. The total complexity for a set of N sequences can be shown to be $O(dnN + \min(u, dn)N^2)$ for doubles and $O(d^2nN + \min(u, d^2n)N^2)$ for triples, where u is the number of unique features. This can be significantly lower than the exponential complexity of the mismatch kernel $O(k^{m+1}|\Sigma|^m nN + min(u', n)N^2)$, where $u' \leq |\Sigma|^k$, and $k = 5, 6$.

We demonstrate the utility of the new kernel in the sequence classification setting of the protein remote homology prediction task. The analysis is conducted on the benchmark set SCOP 1.59 of the protein sequences widely used in the literature [5, 4, 2]. This dataset includes 7329 sequences (only 2862 are labeled), with the labeled sequences classified into 54 families, resulting in 54 detection problems. In a supervised setting, only labeled sequences participate in experiments. In a semi-supervised setting, unlabeled sequences are used to refine the kernel during training/testing. The performance of the proposed kernels is contrasted to the previously published results [2, 4, 5] under the supervised setting in Table 1, with the accompanying ROC-50 plots in Figure 1. The results indicate that the classifiers learned from the new kernel family significantly outperform, both in accuracy and in computational time, the state-of-the-art approaches. Similar improvements are observed in the semisupervised context of [7]. Table 2 and Figure 2 show that the triple kernel outperforms both the profile [2] and the mismatch neighborhood kernels [7] which are reported to perform best in previous studies.

Topic: learning in biological systems, learning algorithms
Preference: oral/poster

Table 1: Comparison of the performance on the SCOP 1.59 dataset under the supervised setting.

Method	ROC	ROC50	# dim.	Time (s)
(5, 1)-mismatch	0.8749	0.4167	3200000	938
SVM-pairwise	0.8930	0.4340	-	-
Fisher	0.7730	0.2500	-	-
(1,5) double	0.8901	0.4629	2000	54
(1,3) triple	0.9148	0.5118	72000	112

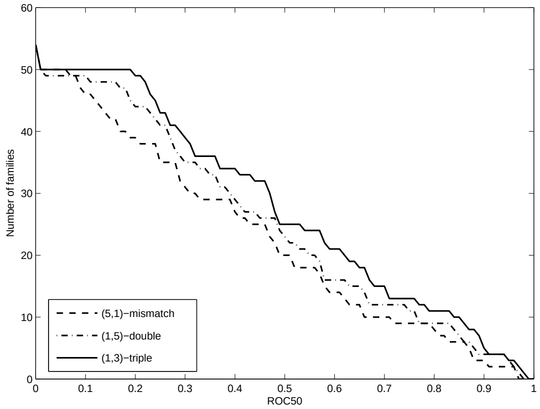

Figure 1: Comparison of the performance (ROC50) in a supervised setting. Spatial kernels (triples and doubles) outperform other supervised methods.

Table 2: Comparison of the performance under the semi-supervised setting.

Method	ROC	ROC50
(5, 1)-mismatch neighborhood	0.9093	0.6745
(5,7.5)-profile	0.9190	0.6069
(1,5)-double	0.9131	0.6279
(1,3)-triple	0.9207	0.7273

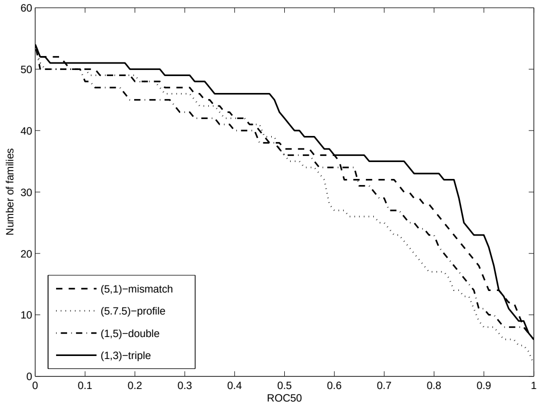

Figure 2: Comparison of the performance (ROC50) in a semi-supervised setting. Spatial triple kernel outperforms both profile and mismatch neighborhood kernels.

The new class of spatially-constrained sparse kernels is expected to scale well and work with the large unlabeled datasets (e.g. nonredundant datasets of protein sequences) in many other challenging semi-supervised sequence classification settings.

References

[1] Jianlin Cheng and Pierre Baldi. A machine learning information retrieval approach to protein fold recognition. *Bioinformatics*, 22(12):1456–1463, June 2006.

[2] Rui Kuang, Eugene Ie, Ke Wang, Kai Wang, Mahira Siddiqi, Yoav Freund, and Christina Leslie. Profile-based string kernels for remote homology detection and motif extraction. In *CSB '04: Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB'04)*, pages 152–160, August 2004. http://www.cs.columbia.edu/compbio/profile-kernel.

[3] Christina Leslie and Rui Kuang. Fast string kernels using inexact matching for protein sequences. *J. Mach. Learn. Res.*, 5:1435–1455, 2004.

[4] Christina S. Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch string kernels for svm protein classification. In *NIPS*, pages 1417–1424, 2002.

[5] Li Liao and William Stafford Noble. Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships. *Journal of Computational Biology*, 10(6):857–868, December 2003.

[6] Sören Sonnenburg, Gunnar Rätsch, and Bernhard Schölkopf. Large scale genomic sequence svm classifiers. In *ICML '05: Proceedings of the 22nd international conference on Machine learning*, pages 848–855, New York, NY, USA, 2005. ACM Press.

[7] Jason Weston, Christina S. Leslie, Dengyong Zhou, André Elisseeff, and William Stafford Noble. Semi-supervised protein classification using cluster kernels. In *NIPS*, 2003.