Efficient Sequence Classification with Spatial Representations

Pavel P. Kuksa, Vladimir Pavlovic

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

Analysis of large-scale sequential data has become an important task in machine learning and pattern recognition, inspired in part by numerous scientific and technological applications such as the document and text classification or the analysis of music data, or biological sequences. We present a general, simple feature representation of sequences, *spatial* representation, that allows efficient inexact matching, comparison and classification of sequential data. This approach, recently introduced for the problem of biological sequence classification, exploits a novel *multi-scale representation* of strings.

We demonstrate that the developed approach is applicable to modeling of sequences in a wide range of sequence domains, both *discrete-* and *continuous*-valued. Experiments using the new features and algorithms on *text document categorization*, *music genre* and *artist recognition* show excellent predictive performance, while demonstrating significant improvements in running time over the existing state-of-the-art sequence classification methods on these large alphabet, large sequence datasets.

Background. A number of state-of-the-art approaches to classification of sequences over finite alphabet Σ rely on measuring sequence similarity using fixed-length representations $\Phi(X)$ of sequences as the *spectra* ($|\Sigma|^k$-dimensional histogram) of counts of short substrings (k-mers), contained, possibly with up to m mismatches, in a sequence, c.f., spectrum/mismatch methods [3, 4, 2]. However, computing similarity scores, or kernels, $K(X,Y)=\Phi(X)^T\Phi(Y)$ using these representations can be challenging, e.g., efficient $O(k^{m+1}|\Sigma|^m(|X| + |Y|))$ trie-based mismatch kernel algorithm [4] strongly depends on the alphabet size and the number of mismatches m. On the other hand, the gapped [3] and subsequence [6] kernels have complexity independent of $|\Sigma|$, but quadratic in the sequence length (subsequence method) or show suboptimal performance compared to other methods (e.g., mismatch).

Method. Similarity evaluation under *spatial representation* amounts to sampling the sequence features at different resolutions and comparing the resulting spectra; similar sequences will have similar spectra at one or more resolutions. Each sampled spatial feature consists of t substrings of length k, with each substring no more than d positions away from its neighbors. We illustrate the spatial features in Figure 1(a). The upper panel shows a typical contiguous spectrum 6-mer and the lower panel shows how a spatial sample method with $k=2,t=3$ would extract features from the string. Much like the spectrum features, the spatial feature "AR__ND_CQ" shown has the value proportional to the number of times it occurs in string X.

While spectrum/mismatch representations rely on *contiguous* string fragments of length k, the *spatial* representation, on the other hand, is *multi-dimensional*, made of variably distanced string fragment combinations (Figure 1(b)). In the figure, we show a spatial embedding ($t = 2$) with string fragments as single symbols ($k=1$) displaced by d (e.g., row "AA" and column $d = 2$ shows number of occurrences of "A__A").
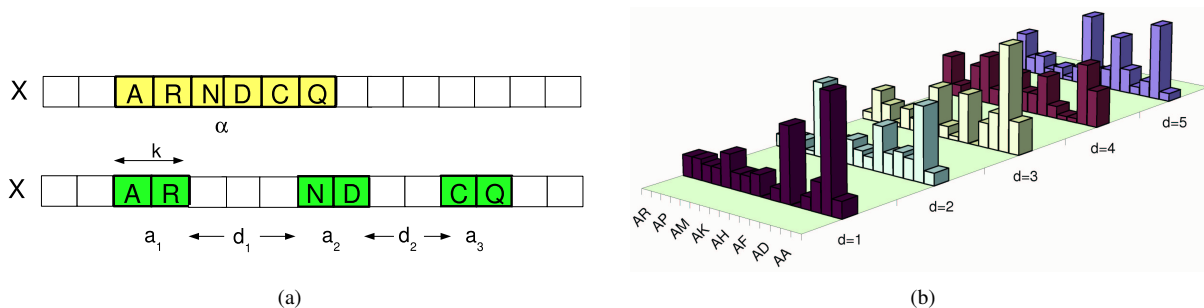

(a)   (b)

Figure 1: Left: Contiguous k-mer feature α of a traditional spectrum feature (top) contrasted with the spatial samples (bottom). Right: The multi-dimensional spatial sample embedding.

Results. We test proposed methods on four distinct multi-class sequence classification tasks: (1) text document categorization, (2) music genre classification, (3) artist identification, and (4) multi-class protein fold prediction.

Text Classification. As shown in Table 1, on a widely used benchmark Reuters dataset (word alphabet $|\Sigma| = 29,224$), *spatial representation*-based kernels (double(1,5), i.e. doubles of words displaced by up to 5 other words) improves over

Topic: data mining, pattern recognition, sequence modeling
Preference: oral/poster

Table 1: Test F1 scores on top Reuters categories. Spatial features improve over baseline and state-of-the-art methods.

Class	TF-IDF	KSG	*Double*	SS-4[†]	NG-4[‡]
Earn	98.70	98.3	**98.76**	97.0	98.40
Acq	97.11	96.8	**97.68**	88.0	93.20
Money	77.61	**84.0**	83.71	76.0	75.70
Grain	93.29	92.5	**93.38**	84.0	84.0
Crude	87.71	89.4	**90.51**	84.0	84.80
Trade	84.26	90.2	**91.23**	73.0	77.90
Interest	71.80	**81.5**	81.15	66.0	71.90
Wheat	80.00	**81.8**	80.92	79.7	79.0
Ship	72.97	81.9	**84.39**	65.0	62.60
Corn	81.63	**87.1**	83.33	63.0	61.0
Macro-average	84.51	88.3	**88.51**	77.57	78.80
Micro-average	93.18	93.9	**94.39**	-	-

KSG=key-substring-group features [10]

[†] *approximate* subsequence kernel [6], [‡] N-gram character kernel [6]

Table 2: Music genre classification.

#	Genre	DWCH[†]	*Double*(1,5)	gapped(4,2)
1	Blues	95.49 (1.27)	93.6 (4.77)	93.8 (2.33)
2	Classical	98.89 (1.1)	95.6 (1.35)	97.2 (1.04)
3	Country	94.29 (2.49)	**94.3 (2.21)**	91.7 (2.02)
4	Disco	92.69 (2.54)	**94.3 (1.41)**	91.9 (1.14)
5	Jazz	97.9 (0.99)	95.5 (2.27)	93.4 (1.29)
6	Metal	95.29 (2.18)	94.7 (1.42)	94.0 (2.37)
7	Pop	95.8 (1.69)	**96.2 (1.75)**	95.5 (1.32)
8	Hiphop	96.49 (1.28)	**97.1 (0.99)**	94.8 (1.25)
9	Reggae	92.3 (2.49)	**95.5 (1.58)**	92.3 (2.02)
10	Rock	91.29 (2.96)	**95.1 (1.66)**	91.7 (1.52)
	Mean	95.04	**95.19**	93.63

[†]: DWCH=Daubechies Wavelet Coefficient Histograms [5]

other state-of-the-art kernel methods, including n-gram, subsequence kernels, TF-IDF word kernels, KSG [10]). KSG [10] often displays performance similar to our approach, but uses significantly more intricate grouping of substrings into key groups (based on, e.g., the maximum parent-child conditional probability in suffix tree decomposition, etc.) Furthermore, it is important to note that even for the large alphabet set ($|\Sigma|$=29, 224) a 8986-by-8986 document similarity matrix for the double(1,5) kernel takes only 16 seconds to compute on a 2.8GHz CPU.

Music Genre Classification. Music genre recognition is a particularly interesting problem in our setting because music data is originally continuous-valued and string representations (in the absence of musical notation) may require a large alphabet. On a standard benchmark dataset [5] (10 genres, each 100 audio sequences, quantized into strings with $|\Sigma| = 1024$) *spatial* kernel achieves better overall performance (Table 2, 10-fold cross-validation) compared to the subsequence/gapped(4,2) kernel, and the DWCHs method [5], an approach specifically developed for music classification. This is achieved using very simple MFCC features that capture only *local* information in the music signals. In contrast, the DWCH method uses more sophisticated features with both *local and global* information. Compared to the gapped kernel with no spatial information, our method achieves better performance in eight out of ten genres. Both of these facts point to importance of considering longer-term spatial relationships in music signals for genre prediction. Similar conclusion carries over to a multi-class setting, Table 3, The raw MFCC features achieve 41.6±3.31 error rate [5]. Our double kernel that incorporates *longer*-term dependency (using d=5) improves the error significantly to 29.2 ± 1.61.

Artist recognition. We also illustrate the utility of our generic spatial string features on multi-class artist identification on the standard *artist20* dataset[1] with 20 artists, 6 albums each (1413 tracks total). Table 4 lists results for 6-fold album-wise cross-validation with one album per artist held out for testing. Using spatial information with quantized MFCC features ($|\Sigma| = 1024$) yields 32.5% error compared to 44% using MFCC features alone [1], indicating that our spatial features may be well suited for this task, especially when coupled with more domain-specific information.

Table 3: Multi-class music genre classification

method	Error	Top-2 Error	F1	Top-2 F1
gapped	34.5±2.6	19.9±2.27	65.35	80.31
double	29.2±1.61	17.5±1.77	70.82	82.62
triple	29.3±1.86	17.3±1.89	70.61	82.78

Table 4: Artist recognition performance

method	Error	Top-2 Error	F1	Top-2 F1
gapped	44.66	32.24	55.33	67.99
double	**32.50**	**21.51**	**67.56**	**78.63**
triple	32.69	21.15	67.43	78.67

Table 5: Running time (s) for kernel computation between two strings on *real* data

	protein (semi-sup)	text	music
n, $|\Sigma|$	36672, 20	242, 29224	6892, 1024
(5,1)-mismatch	1.6268	20398	526.8
subseq. (p=3)	1222.4	0.4846	2.4321
Spatial kernel (triple)	**0.1967**	**7.5e-03**	**3.45e-02**

*n-sequence length, $|\Sigma|$-alphabet size

Running time. One important benefit of our approach lies in the computational efficiency of evaluating similarity of sequences (kernel values). As shown in Table 5, both double and triple kernels demonstrate order of magnitude running time improvements over other algorithms.

References

[1] Dan Ellis. Classifying music audio with timbral and chroma features. In *Proc. Int. Conf. on Music Information Retrieval ISMIR-07*, 2007.

[2] Rui Kuang, Eugene Ie, Ke Wang, Kai Wang, Mahira Siddiqi, Yoav Freund, and Christina S. Leslie. Profile-based string kernels for remote homology detection and motif extraction. In *CSB*, pages 152–160, 2004.

[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] Tao Li, Mitsunori Ogihara, and Qi Li. A comparative study on content-based music genre classification. In *SIGIR '03*, pages 282–289, New York, NY, USA, 2003. ACM.

[6] Huma Lodhi, Craig Saunders, John Shawe-Taylor, Nello Cristianini, and Chris Watkins. Text classification using string kernels. *J. Mach. Learn. Res.*, 2:419–444, 2002.

[7] Iain Melvin, Eugene Ie, Jason Weston, William Stafford Noble, and Christina Leslie. Multi-class protein classification using adaptive codes. *J. Mach. Learn. Res.*, 8:1557–1581, 2007.

[8] B. Reva, A. Kister, S. Topiol, and I. Gelfand. Determining the roles of different chain fragments in recognition of immunoglobulin fold. *Protein Eng.*, 15(1):13–19, 2002.

[9] Jason Weston, Christina Leslie, Eugene Ie, Dengyong Zhou, Andre Elisseeff, and William Stafford Noble. Semi-supervised protein classification using cluster kernels. *Bioinformatics*, 21(15):3241–3247, 2005.

[10] Dell Zhang and Wee Sun Lee. Extracting key-substring-group features for text classification. In *KDD '06*, pages 474–483, New York, NY, USA, 2006. ACM.

[1]http://labrosa.ee.columbia.edu/projects/artistid/