2D similarity kernels and representations for sequence data

Pavel P. Kuksa

NEC Laboratories America, Inc
pkuksa@nec-labs.com

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.

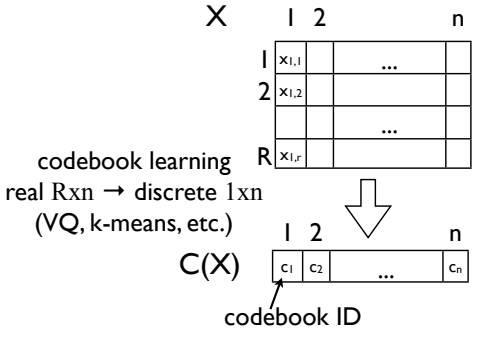
In this work, we consider general, simple 2D *matrix representations* of sequences, and introduce a class of *2D similarity kernels* that allows efficient inexact matching, comparison and classification of sequence inputs in the form of *sequences of R-dim. feature vectors*. The developed approach is applicable to a wide range of sequence domains, both *discrete*- and *continuous*-valued, such as *music*, *images*, or *biological sequences*. Experiments using the new 2D representations and kernels on *music genre* and *artist recognition* show excellent predictive performance with significant *25%-40%* improvements over the existing state-of-the-art sequence classification methods.

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]. This essentially amounts to analysis of 1D sequences over finite alphabets Σ with 1D k-mers as basic sequence features. However, original input sequences are often in the form of *sequences of feature vectors*, i.e. each input sequence X is a *sequence of R-dim. feature vectors* which could be considered as $R \times |X|$ feature matrix. Examples of these include

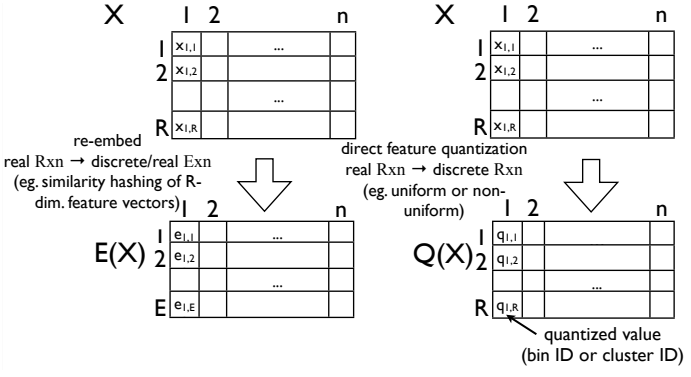
- *Music data*. Each music sequence X in commonly used MFCC representation is a sequence of 13-dim. MFCC features (2D sequence of size $13 \times |X|$).
- *Image data*. Each image could be considered as a 2D sequence of feature vectors corresponding to decomposition of an image into a regular grid of smaller image blocks (e.g., as in [6]);
- *Biological data*. Protein sequences could be considered as 2D sequences of R-dim. feature vectors describing physical/chemical properties of individual amino acids.

In this work, we aim at methods that directly exploit these richer 2D sequence representations to improve accuracy and propose a family of 2D similarity kernels that as we show empirically provide effective improvements in practice over traditional 1D sequence kernels for a number of challenging classification problems.

2D sequence representations. In a typically used *codebook learning* framework, input sequences (or sets) of features vectors are typically first encoded using codebook IDs (Figure 1a), then standard 1D string kernel methods can be applied. In this work we consider alternative *2D representations* (Fig. 1b) of *sequences of feature vectors*. In particular, we consider two approaches: (1) encoding original continuous-valued feature vectors in discrete (binary) space using e.g. similarity hashing approach [7] (Figure 1b); (2) directly quantizing each feature using, for example, uniform binning (Figure 1b). We will show in the experiments that using these *2D representations* can *significantly improve* predictive accuracy compared to traditional 1D kernel representations as well as other state-of-the-art approaches.



(a) Typical representation for sequence X: R-dim. input feature vectors encoded using corresponding codebook IDs. 1D string kernel is used to measure similarity between sequences.



(b) Proposed approach. Input sequence X of R-dim feature vectors is represented in 2D using direct feature quantization $Q(X)$ or embedding $E(X)$. 2D string kernel is used to measure sequence similarities.

Figure 1: Proposed *2D representations* contrasted with traditional codebook-based approach and 1D string kernels.

2D similarity kernels. Similarity evaluation between two 2D sequences X and Y under 2D *matrix representation* amounts to comparing pairs of 2D submatrices contained in X and Y. A 2D string kernel can be defined for 2D sequences X and Y as

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

Table 1: Music genre recognition (10-class, 13-dim. MFCC features only)

method	Error, %	F1
Baseline 1 (non-MFCC): DWCH [5]	21.5	-
Baseline 2 (best): AdaBoost (MFCC,FFT,LPC,etc)	17.5	-
Vector quantization (discrete 1D)		
Spectrum (k=1)	34.5	65.61
Mismatch-(k=5,m=2)	32.6	67.51
SSSK-(k=1,t=2,d=5)	31.1	69.08
Manifold Spectrum-(k=1)	26.9	73.31
Manifold SSSK-(k=1,t=2,d=5)	25.0	75.27
Similarity hashing (binary 2D)		
Spectrum-(k=8), B=64	27.8	72.25
Manifold Spectrum-(k=8)	24.1	76.29
Uniform direct feature quantization (discrete 2D)		
Spectrum-(k=1)	28.5	72.06
SSSK-(k=2,t=2,d=5)	26.3	73.88
Manifold Spectrum-(k=1)	23.0	77.23
Manifold SSSK-(k=2,t=3,d=5)	**17.2**	**83.09**
Manifold SSSK-(k=2,t=3,d=5)+FFT	**14.1**	**86.14**

Table 2: Music genre recognition (ISMIR contest, 6-class, 13-dim. MFCC features only)

method	Error, %	F1
Vector quantization (discrete 1D)		
Spectrum (k=1)	24.15	68.99
Manifold Spectrum-(k=1)	19.62	76.17
Similarity hashing (binary 2D)		
Spectrum-(k=6), B=32	22.63	71.57
Manifold Spectrum-(k=6)	17.83	79.06
Uniform direct feature quantization (discrete 2D)		
Manifold Spectrum-(k=3)	**16.74**	**80.57**

Table 3: Music artist identification (20-class, 13-dim. MFCC features only)

method	Error, %	F1
Vector quantization (discrete 1D)		
Spectrum (k=1)	42.97	57.26
Similarity hashing (binary 2D)		
Manifold Spectrum-(k=6)	34.62	66.22
Uniform direct feature quantization (discrete 2D)		
Manifold Spectrum-(k=6)	**25.67**	**74.79**

$$K_{2D}(X,Y) = \sum_{\alpha_{2D} \in X} \sum_{\beta_{2D} \in Y} \mathcal{K}(\alpha_{2D}, \beta_{2D}) \quad (1)$$

where α_{2D} and β_{2D} are $|R| \times k$ submatrices of X and Y and $\mathcal{K}(\alpha_{2D}, \beta_{2D})$ is a kernel function defined for measuring similarity between two submatrices. One possible definition for $\mathcal{K}(\cdot,\cdot)$ that we use in this work is row-based similarity

$$\mathcal{K}(\alpha_{2D}, \beta_{2D}) = \sum_{i=1}^{R} I(\alpha_{2D}^i, \beta_{2D}^i) \quad (2)$$

where $I(\cdot,\cdot)$ is a similarity/indicator function for matching 1D rows α_{2D}^i and β_{2D}^i. Using Eq. 2, 2D kernel in Eq. 1 can be written as $K_{2D}(X,Y) = \sum_{i=1}^{R} \sum_{\alpha_{2D} \in X} \sum_{\beta_{2D} \in Y} I(\alpha_{2D}^i, \beta_{2D}^i)$ which can be efficiently computed by running spectrum kernel with 1D k-mer matching function $I(\cdot,\cdot)$ B times, i.e. for each row $b = 1 \ldots B$.

Results. We test proposed methods on a number of multi-class sequence classification tasks: (1) 10-class music genre classification[1], (2) 6-class music genre recognition (ISMIR contest[2]), and (3) 20-class music artist identification (artist20 dataset[3]). For all tasks input sequences are sequences of 13-dim. MFCCs, and we test traditional 1D vector quantization (VQ) approach, and two proposed 2D representations using *similarity hashing* (2D) and *direct uniform feature quantization* (2D) (Fig. 1b). For all kernels we test with and without embedding resulting kernel feature representations into multinomial manifold. We use state-of-the-art spectrum/mismatch [3] and spatial (SSSK) [2] kernels as our basic 1D similarity kernels.

Music genre recogntion. As shown in Table 1, on a widely used benchmark dataset [5] (10 genres, each with 100 sequences), *2D k-mer*-based kernels improve over traditional 1D kernels and other state-of-the-art methods, including DWCH [5], aggregate feature AdaBoost [1], approaches specifically developed for music classification that also use many other features in addition to MFCC. For example, using 2D similarity kernels with SSSK [2] achieves significantly higher accuracy of 82.8% compared to only 75.0% when using 1D kernels. We observe similar overall improvements for 2D similarity kernels on another benchmark dataset (ISMIR genre contest), Table 2.

Artist recognition. We also illustrate the utility of our 2D string kernels on multi-class artist identification on the standard *artist20* dataset with 20 artists, 6 albums each (1413 tracks total). Table 3 lists results for 6-fold album-wise cross-validation with one album per artist held out for testing. Using 2D similarity kernels with direct uniform quantization of MFCC features yields a much lower 25.7% error compared to 42.9% for 1D kernels.

References

[1] James Bergstra, Norman Casagrande, Dumitru Erhan, Douglas Eck, and Balázs Kégl. Aggregate features and adaboost for music classification. *Mach. Learn.*, 65:473–484, December 2006.

[2] Pavel P. Kuksa and Vladimir Pavlovic. Spatial representation for efficient sequence classification. In *ICPR*, 2010.

[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] Zhiwu Lu and H.H.S. Ip. Image categorization with spatial mismatch kernels. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:397–404, 2009.

[7] Yair Weiss, Antonio Torralba, and Rob Fergus. Spectral hashing. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 1753–1760. 2009.

[1] http://opihi.cs.uvic.ca/sound/genres

[2] http://ismir2004.ismir.net/genre_contest/index.htm

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