Vol.2012-BIO-31 No.4
2012/10/15

不連続な部分構造に基づく文字列カーネルとそのタンパク質の分類への応用

小野寺 拓[1,a)]　渋谷 哲朗[1,b)]

概要：本稿ではサポートベクターマシンによる文字列分類のためのカーネルであるギャップ付きスペクトラムカーネルを提案する。ギャップ付きスペクトラムカーネルは文字列をそこに含まれる不連続なパターンによって特徴づける。このカーネルは基本的な文字列カーネルであるスペクトラムカーネルと同様、文字列をそれに含まれる k-mer により特徴づける。スペクトラムカーネルは文字列の長さ n に対して線形時間で計算可能である反面、より高い学習能力を持つその変種の計算には一般に $\Omega(p(k)n)$ の計算量が必要である。ただしここで $p(\cdot)$ は高次の多項式である。我々は $O(gn)$ 時間でのギャップ付きスペクトラムカーネルの計算法を与える。ただし g は k 以下の定数である。また、タンパク質の分類実験の結果から、このカーネルがスペクトラムカーネルの変種である wild card kernel と同等の学習能力をより短時間に実現できることを示す。

1. 導入

1.1 背景

近年サポートベクターマシンとカーネル法を用いた学習手法がテキスト分類や音声認識等の様々な用途に応用されている。この方法の最大のポイントは特徴ベクトルを明示的に保持することなくカーネル函数、すなわち特徴ベクトル同士の内積を効率的に計算できれば高速な計算と高い学習能力を実現できる点である。そのため対象や用途に応じて様々な特徴写像およびそれに応じたカーネル函数の効率的な計算アルゴリズムが考えられてきた。とくに文字列はそれ自体が普遍的なデータであるのはもちろん、その他の離散的データにも木など文字列として自然に表現できるものが多く存在するためカーネル法に基づく学習手法の中でも文字列を対象としたものは最重要なものの一つといえる。

1.2 関連研究

最も基本的な文字列カーネルであるスペクトラムカーネル（spectrum kernel）[1] では文字列中の特定の長さ k の部分文字列の出現回数をもとに特徴ベクトルが定義され、入力文字列 x と y に対するカーネル函数の値を $\Theta(n)$ の時間で計算可能である。ただしここで n は x, y の長さの合計を表す。スペクトラムカーネルを元に mismatch kernel [2], restricted gappy kernel, substitution kernel, wildcard kernel [3] 等の様々な文字列カーネルが考案されてきた。これらのカーネルはスペクトラムカーネル同様長さ k の部分文字列の出現回数に依るが、部分文字列そのもののみならずそれらからの誤差が一定値以下のものもカウントする点が異なる。ここで誤差の定義の仕方はカーネルごとに異なる。これらのカーネルはスペクトラムカーネルより高い学習性能を実現するが、一方で w を誤差の閾値とするとカーネル函数の計算に $\Omega(k^w n)$ の時間を必要とする。

1.3 我々の貢献

我々はスペクトラムカーネルの自然な一般化であるギャップ付きスペクトラムカーネルを提案する。このカーネルは前節で挙げたカーネルと同様に部分文字列の出現回数に基づくが、k 以外のパラメータとしてギャップパターンをもつ。対応するカーネル函数は b-接尾辞配列及び b-高さ配列 [5] を用いて $O(gn)$ 時間で計算できる。ただしここで g はギャップパターンにより決まる k 以下の定数である。タンパク質の分類実験の結果は、少数のギャップパターンに対するギャップ付きスペクトラムカーネルを組み合わせることで wild card kernel 等の複雑なカーネルと同等の学習性能をスペクトラムカーネルと同等の計算速度で実現できることを示す。

2. ギャップ付きスペクトラムカーネル

文字集合 Σ は有限な全順序集合とする。とくに断りがな

[1] Human Genome Center, University of Tokyo, 4-6-1, Shirokanedai, Minato-ku, Tokyo 108-8639, JAPAN
[a)] tk-ono@hgc.jp
[b)] tshibuya@hgc.jp

い限り文字列とは $\Sigma^* := \cup_{i=0}^{\infty}\Sigma^i$ の元であるとする。文字列 x の長さを $|x|$ で表す。

以降では $\{1,2,\ldots,\sigma^k\}$ と k-mer の間のある全単射が与えられていると仮定し、i に対応する k-mer を i 番目の k-mer とよぶ。k と m を $1 \leq k \leq m$ を満たす整数とする。b を長さ m で k 個の 1 を含む二進の文字列とし、i_j を b 中の j 番目の 1 の位置とする。$mask_b$ を $\Sigma^m \ni x_1x_2\ldots x_m \mapsto x_{i_1}x_{i_2}\ldots x_{i_k} \in \Sigma^k$ なる写像とした場合、ギャップ付きスペクトラムカーネルとは次の特徴写像により決まるカーネルである:

$$\phi_b : \Sigma^* \ni x \mapsto \sum_{i=1}^{|x|-m+1} \mathbf{1}_{mask_b(x_ix_{i+1}\ldots x_{i+m-1})}.$$

ただし $\mathbf{1}_s$ は i 番目の k-mer が s である場合、またその場合に限り i 番目のエントリが 1 であり、それ以外の場合は 0 であるようなベクトルである。

対応するカーネル関数 $K : (x,y) \mapsto \phi_b(x)\phi_b(y)$ は b-接尾辞配列および b-高さ配列 [5] というデータ構造を用いて効率的に計算できる。文字列 S の i 番目の接尾辞 $S[i]S[i+1]\ldots S[|S|]$ を S_i と表すとしたとき、S に対する b-接尾辞配列 b-SA とは $\{i\}_{i=1}^n$ の要素を $mask_b(S_i)$ の辞書式順序で並べたものを格納した配列であり、b-高さ配列 b-Hgt とは $mask_b(S_{b\text{-}SA[i]})$ と $mask_b(S_{b\text{-}SA[i-1]})$ の最長共通接頭辞長を i 番目に持つ配列である。$K(x,y)$ の計算は、x と y を連結した文字列に対する b-SA 及び b-Hgt の構築とその走査に帰着し、とくに計算時間は前者に依存する。これらの配列はともに $O(gn)$ 時間で構築可能であるただしここで $n := |x|+|y|$ でありまた、g は b 中の連続する 1 の個数でとくに $g \leq k$ を満たす。

上の定義で任意の $1 \leq i \leq m$ に対して $b[i] = 1$ である場合は m-スペクトラムカーネルと一致する。また、m を固定して $m-k$ が w 以下であるような全ての b を列挙しつつ、対応するカーネル関数を足して得られるカーネルは $(m,w,1)$-wildcard kernel[3] に相当する。

3. 実験

ギャップ付きスペクトラムカーネルを用いてタンパク質の分類を行った。データは既存の文字列カーネルの研究でも広く使われてきたタンパク質データベースである SCOP[4] から取得した。SCOP は階層的なデータベースで、各レベルには Class や Family といった名前がつけられている。ここでは 50 個以上の配列を含む 13 個の Family に対して、与えられたタンパク質が各 Family の元かどうかを予測した。まず全データをランダムに二分割してテストデータと訓練データを用意した。次に長さ 5 以下で 0 を高々 2 個含むような 16 個のギャップパターン全てを列挙しながら対応するギャップ付きスペクトラムカーネルのカーネル行列の部分和をとった。これから $(5,2,1)$-wild card kernel に収束するカーネルの系列が得られるが、これらを用いてテストデータの分類を行った。列挙の順番はランダムな順番と訓練データ内の 2fold での交差検定における性能により整列した順番を用いた。また比較のため $2 \leq k \leq 4$ に対する k-スペクトラムカーネルを用いた分類も行った。訓練データに対する交差検定、テストにおけるカーネルの性能の評価は全て ROC 曲線下の面積である AUC (area under curve) を用いた。

図 1 は Family b.1.1.1 に対する実験結果である。赤い線が訓練データを用いて選択した系列、各黒い線がランダムな系列、水平な線は $2 \leq k \leq 4$ に対する k-スペクトラムカーネルのなかで最もスコアの高かったものをそれぞれ表す。その他の Family に対する実験結果は付録に示す。

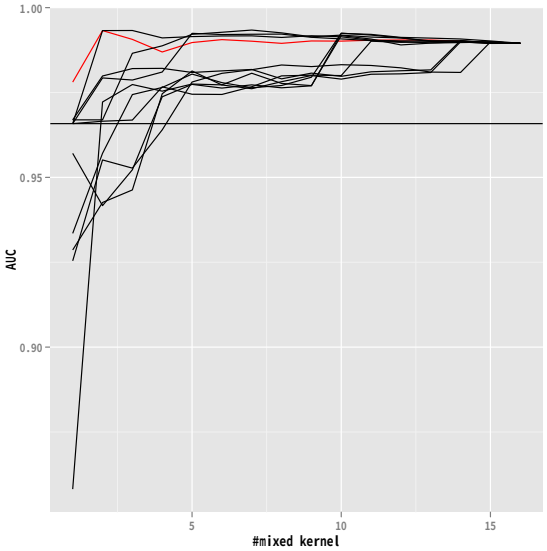


図 1 Family b.1.1.1 に対する実験結果

組み合わせるカーネルの個数を増やすと AUC スコアは増える傾向があるが、実際には少数のカーネルを組み合わせた時点でスコアは wild card kernel のスコアとほぼ等しくなり以降は殆ど変化しない場合が多い。とくに b 以下の Class に属する Family に対してはこの傾向が顕著である。このような Family に対してはより少ない計算量で wild card kernel と同等の精度での分類が行える。とくにテスト時に $O(1)$ 個のカーネルの組み合わせを用いる場合の計算量は $O(gn)$ である。

参考文献

[1] Christina S. Leslie, Eleazar Eskin, and William Stafford Noble. The spectrum kernel: A string kernel for svm protein classification. In *Pacific Symposium on Biocomputing*, pages 566–575, 2002.

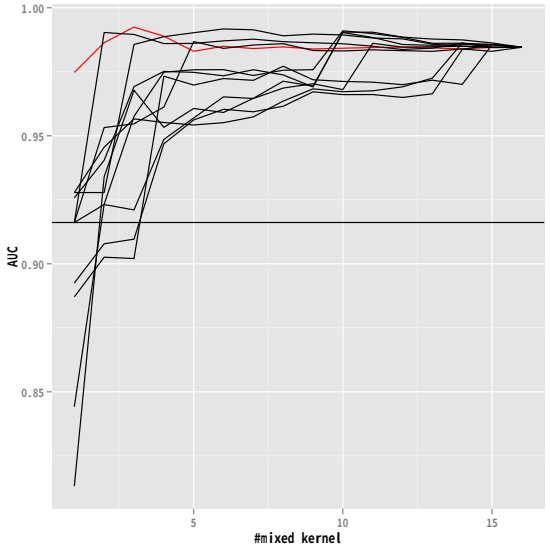
[2] Christina S. Leslie, Eleazar Eskin, Jason Weston, and William Stafford Noble. Mismatch string kernels for svm protein classification. In Suzanna Becker, Sebastian Thrun, and Klaus Obermayer, editors, *NIPS*, pages 1417–1424. MIT Press, 2002.

[3] Christina S. Leslie and Rui Kuang. Fast string kernels using inexact matching for protein sequences. *Journal of Machine Learning Research*, 5:1435–1455, 2004.

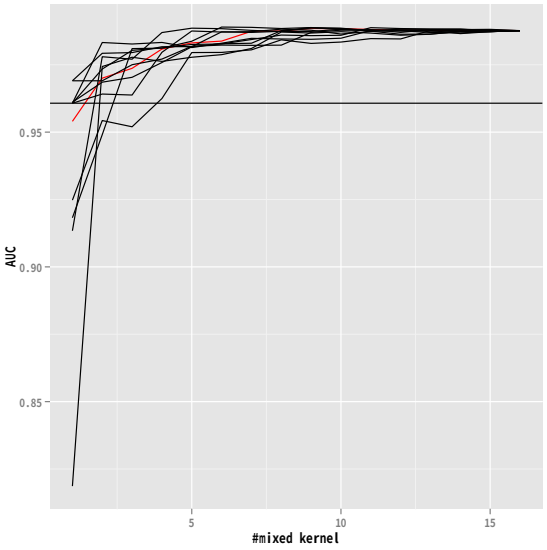
[4] Alexey G. Murzin, Steven E. Brenner, Tim Hubbard, and Cyrus Chothia. Scop: A structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536 – 540, 1995.

[5] Taku Onodera and Tetsuo Shibuya. An index structure for spaced seed search. In Takao Asano, Shin-Ichi Nakano, Yoshio Okamoto, and Osamu Watanabe, editors, *ISAAC*, volume 7074 of *Lecture Notes in Computer Science*, pages 764–772. Springer, 2011.

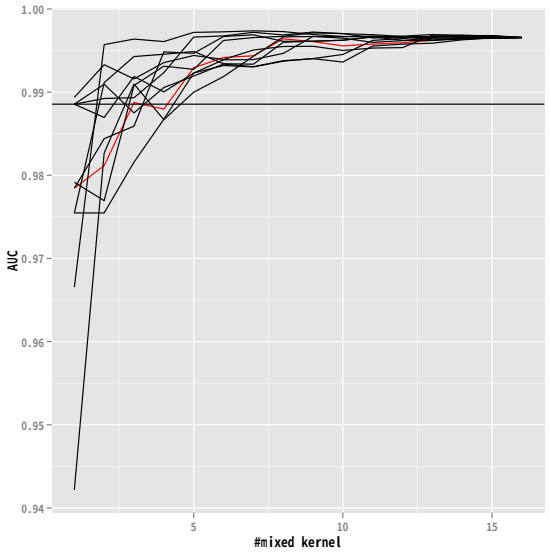
付　　録



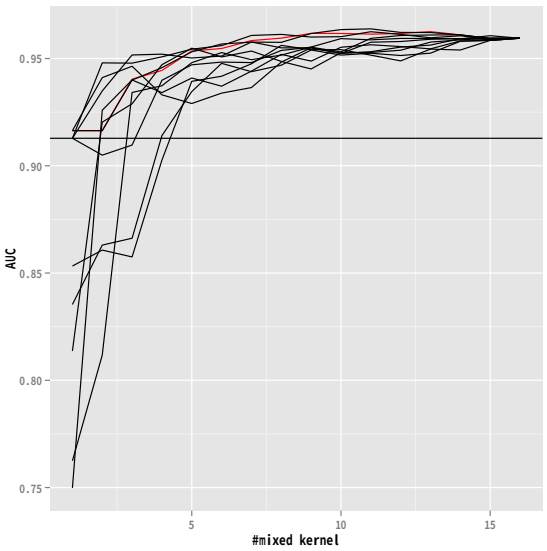
(a) b.1.1.4



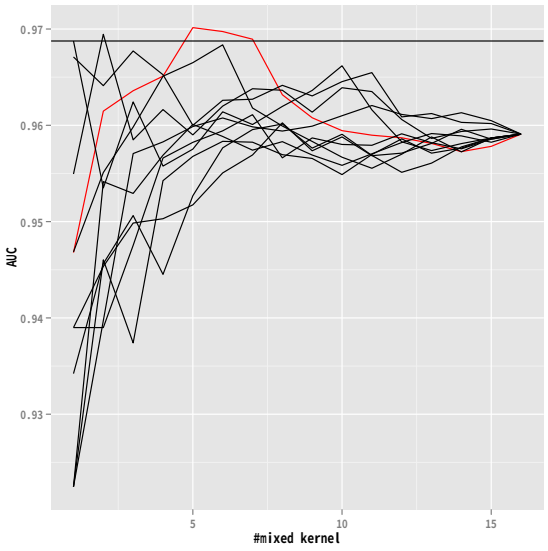
(b) b.1.2.1



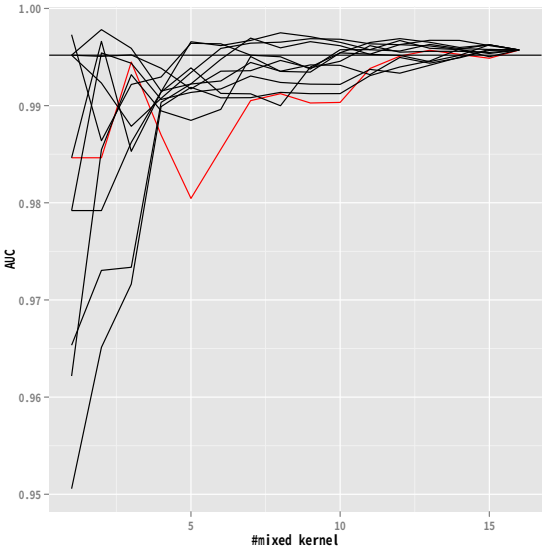
(c) b.36.1.1



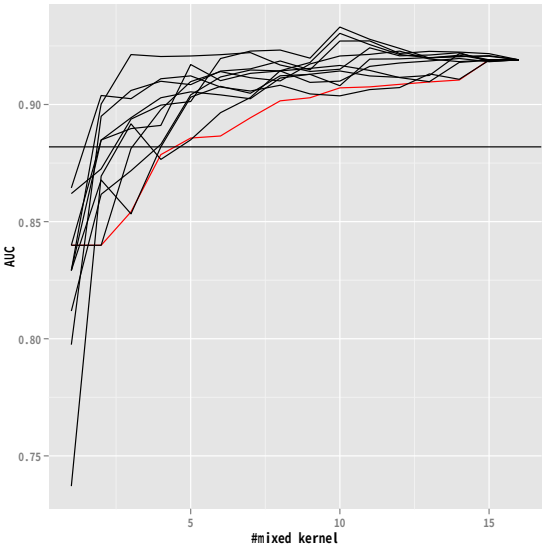
(d) b.40.4.5



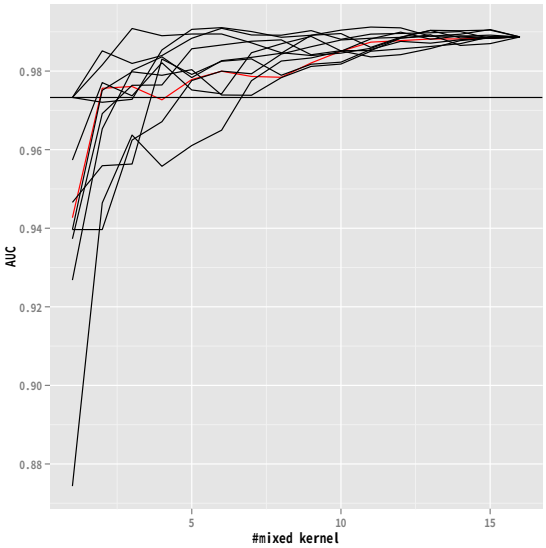
(e) c.2.1.2



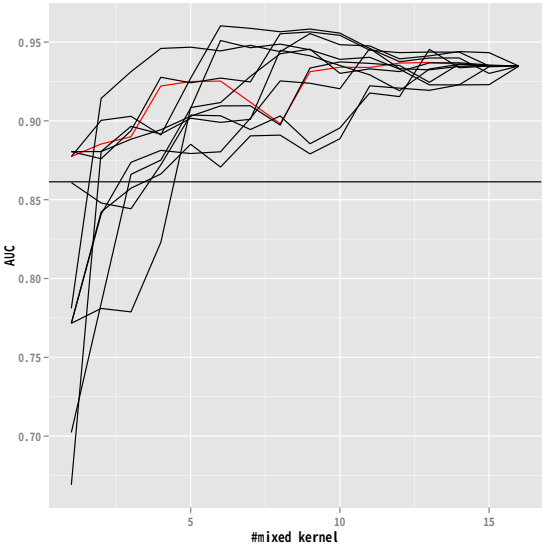
(f) c.37.1.8



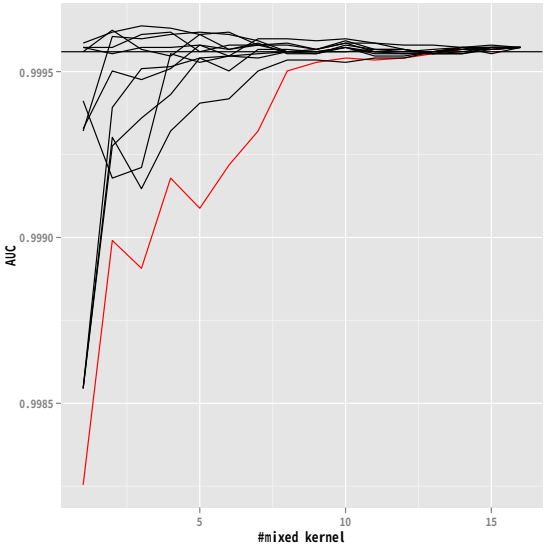
(g) c.94.1.1



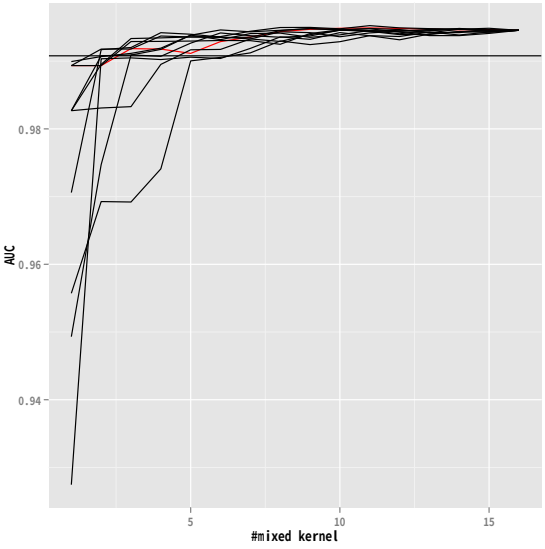
(h) d.58.7.1



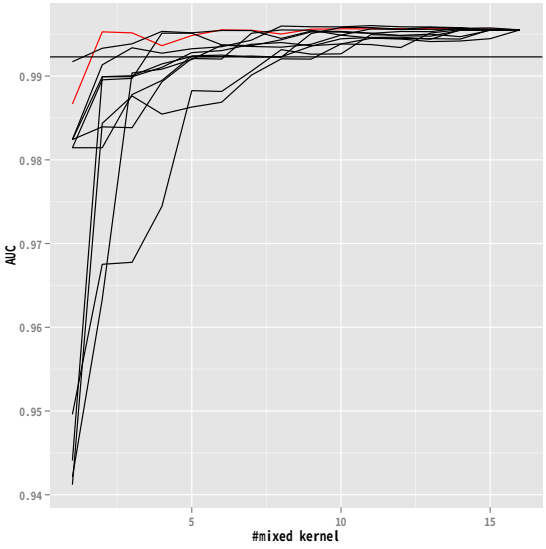
(i) d.108.1.1



(j) d.144.1.7



(k) g.37.1.1



(l) g.39.1.3