D_030

縮約構造とカーネル法を利用した分類手法の検討

A Step towards New Classification Method with Kernel Method and Compressed Structure of Sample Data

大野 博之† 稲積 宏誠†
Hiroyuki Oono Hiroshige Inazumi

1. はじめに

分子生物学において遺伝子機能予測に代表されるゲノム解析が注目されており，塩基配列などの遺伝子情報から遺伝子機能を予測することや分類精度を向上させることが現在の重要な研究課題の1つとなっている．

遺伝子機能予測は，「配列が類似している遺伝子はその機能も類似している」という考え方に基づいている．そのため機能が未知のDNA配列が得られたとき，そのDNA配列がすでに機能の判明しているDNA配列のクラスに分類されるか否かで，そのDNA配列の機能予測を行う方法が多く用いられている．現在では機能の判明しているDNA配列は膨大であるが，配列そのものと比較する必要があり，比較のコストは大きい．

そこで本稿では，機能の判明しているDNA配列から縮約構造を求め，冗長構造を除去することにより，精度向上を実現し，未知の塩基配列の分類を行う手法を検討する．特に，類似度を評価する方法として，塩基配列上の縮約された文字を含む d 次の文字分布を特徴ベクトルとして，分類問題で有効視されているカーネル関数に利用した判別分類を実現し，その適用可能性と有効性を検討する．

2. カーネル法の基本概念

カーネル法 (kernel method) とは一連の機械学習の手法であり，その特徴は，対象領域に対する事前知識をカーネル関数 (kernel function) の形で表現することにある．対象全体の集合を $\boldsymbol{X}$ とすると，カーネル関数は2つの対象 $\boldsymbol{x}, \boldsymbol{x}' \in \boldsymbol{X}$ に対して

$$K : [\boldsymbol{x}, \boldsymbol{x}'] \rightarrow \boldsymbol{R} \tag{1}$$

によって表される関数となる．カーネル関数は2つの対象の「類似度」を表していると考えることができるので，類似度の形を自らの事前知識で表現し，それをカーネル関数として学習法に組み込むことができる．

事前知識を表現するのが特徴空間 (feature space) である．特徴空間をうまく表現することによって，学習が困難な非線形モデルを学習が容易な線形モデルへと変換することができる．そこで，式(2)のような前処理を考える．

$$\boldsymbol{x} \rightarrow \boldsymbol{\phi}(\boldsymbol{x}) \tag{2}$$

$\boldsymbol{\phi}$ は入力空間 $\boldsymbol{X}$ から特徴空間 $\boldsymbol{F}$ への非線形写像であり，$\dim \boldsymbol{F}$ は $\dim \boldsymbol{X}$ よりも遙かに大きいとする．ここで k 次までの全ての項を含む特徴空間に展開することを考えると，その特徴空間の次元数は $\binom{k+\dim \boldsymbol{F}-1}{k}$ となり，このような写像を計算機上で実現するのは計算量やメモリ使用量の問題から難しい．そこで，まず入力空間上での線形モデル $f(\boldsymbol{x})$ の重みを訓練サンプルの線形結合 $\boldsymbol{w} = \sum_{i=1}^{n} \gamma_i \boldsymbol{x}_i$ で表せば，$f(\boldsymbol{x})$ は $\boldsymbol{x}$ と $\boldsymbol{x}_i$ の内積の線形和 $f(\boldsymbol{x}) = \sum_{i=1}^{n} \gamma_i (\boldsymbol{x} \cdot \boldsymbol{x}_i)$ で表される．このような線形モデルを高次元空間で適用すると，

$$f(\boldsymbol{\phi}(\boldsymbol{x})) = \sum_{i=1}^{n} \gamma_i (\boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}_i)) \tag{3}$$

となる．ここで，$\boldsymbol{\phi}$ の内積に対応するカーネル関数 K

$$K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{\phi}(\boldsymbol{x}') \tag{4}$$

の存在を仮定すると，高次元空間での線形モデルは K のみによって書くことができ，

$$f(\boldsymbol{\phi}(\boldsymbol{x})) = \sum_{i=1}^{n} \gamma_i K(\boldsymbol{x}, \boldsymbol{x}_i) \tag{5}$$

となって写像 $\boldsymbol{\phi}$ の計算を回避できる．

カーネル法は上記のカーネル関数を用いて学習を行うことによって判別関数 $f(\boldsymbol{x})$ を構築し，判別分類を行うものである．

3. 縮約構造を利用した分類

塩基配列データを分類するために本提案手法では次の3つのステップが必要となる．

1. 訓練データとなる塩基配列間で，共通する部分配列を抽出し，縮約構造とする．
2. 訓練データから特徴ベクトルを d 次の *count kernel* により定義し，カーネル関数を作成する．
3. 作成したカーネル関数を搭載する判別器を用意し，評価データに対して分類・性能評価を行う．

3.1 縮約構造の抽出

縮約構造の定義と抽出には，さまざまな方法が考えられるが，本稿では，アラインメントによる局所的な類似性を組み合わせることにより，塩基配列群全体の縮約構造を得ることとした．まず，配列群全体に対してアラインメントを行い，得られた系統樹に基いてクラスタリングを行う．次にクラスタごとに再度アラインメントを行い，共通する部分配列を基点とした領域に分割する．そして分割した各領域内で配列の並びが一致するような最大の共通部分配列を抽出する．ただし縮約によって削除された部分配列は1つの不定文字 $*$ に置き換え，何らかの配列があったことを示すものとする．各領域でこのようにして得られた局所的な共通部分配列を組み合わせることにより，そのクラスタの縮約構造とし，同様にしてすべてのクラスタから縮約構造を得る．

3.2 カウントカーネルを利用したカーネル関数

カウントカーネル (count kernel:CK)[1] とは，文字の出現頻度を特徴ベクトルとして用いるカーネルで，DNA塩基 $\{A, C, G, T\}$ の特徴ベクトルは 1st order CK ならば4次元，2nd order CK ならば $\{AA, AC, \cdots, TT\}$ のように16次元で表現できる．例えば 2nd order CK の場合，m 文字の配列 $\boldsymbol{x} = x_1 x_2 x_3 ... x_m$ $(x_i \in \{A, C, G, T\})$ に対して，特徴ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ を以下のように定義できる．

† 青山学院大学 理工学部 情報テクノロジー学科

$$\boldsymbol{\phi}(\boldsymbol{x}) = (c_{aa}(\boldsymbol{x}), c_{ac}(\boldsymbol{x}), \cdots, c_{tt}(\boldsymbol{x})) \quad (6)$$

$$c_{kk'}(\boldsymbol{x}) = \frac{1}{m-1}\sum_{i=1}^{m-1}\delta(x_i,k)\delta(x_{i+1},k')\ ,\ \delta(x_i,k) = \begin{cases}1\ (x_i = k)\\ 0\ (x_i \neq k)\end{cases} \quad (7)$$

すると,CK を利用したカーネル関数は, 以下のような特徴ベクトルの内積で定義される.

$$K_c(\boldsymbol{x},\boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})\cdot\boldsymbol{\phi}(\boldsymbol{x}') = \sum c_{kk'}(\boldsymbol{x})c_{kk'}(\boldsymbol{x}') \quad (8)$$

本提案手法における縮約文字には不定文字 $*$ が含まれており, 特徴ベクトルの次元数, すなわち特徴要素数は DNA 塩基のみを用いる場合に比べ増加するため, 特徴表現が多彩となる. その結果, 計算量は増加してしまうが縮約によって対象文字列の長さ自体は短くなっている.

一方, 訓練データからは縮約構造を得るが, 例外的にクラスタリング結果が 1 データとなり, 実質的に縮約構造が得られない場合もある. 評価データは縮約をせずに取り扱う. このような場合には, 不定文字 $*$ を含む特徴要素の頻度分布は得られない. そこで, 本稿では CK の特徴ベクトル計算を以下のように拡張した.

1. まず不定文字 $*$ を含む d 次の頻度分布を特徴ベクトルとして計算する.

2. 次に, 不定文字 $*$ を含む特徴要素において $*$ をワイルドカードとして捉え, $*$ を含まない他の特徴要素の中から, 任意一致した特徴要素の頻度分布の平均を加算することで更新を行う.

これは, 不定文字 $*$ を含む構造そのものの一致度とあらゆる文字として捉えた時の一致度を共に考慮したいためである.

2nd order CK の場合を例にすると, 特徴ベクトル $\boldsymbol{\phi}(\boldsymbol{x})$ は, 式 (9) のようになり, 不定文字 $*$ を含む特徴要素 c_{a*} は, 式 (10) のように更新される.

$$\boldsymbol{\phi}(\boldsymbol{x}) = (c_{aa}(\boldsymbol{x}), \cdots, c_{a*}(\boldsymbol{x}), \cdots, c_{tt}(\boldsymbol{x}), c_{*a}(\boldsymbol{x}), \cdots) \quad (9)$$

$$c_{a*}(\boldsymbol{x}) = c_{a*}(\boldsymbol{x}) + \frac{1}{4}\sum c_{ai}(\boldsymbol{x}), (i = a,c,g,t) \quad (10)$$

この CK を不定文字 $*$ に関して拡張したものを本稿では CK* と呼ぶこととする.

3.3 判別識別器

判別分類は,3.2 で定義したカーネル関数を Kernel Fisher Discriminant Analysis(KFDA) [2] に適用して行う. パラメータ α_i は,n 個の訓練データによる学習によって決定され, 判別関数 $f(\boldsymbol{x})$ に対する閾値は, 訓練データで得られた値の平均値を算出し, その中間値とする.

$$f(\boldsymbol{x}) = \sum_{i=1}^{n}\alpha_i \boldsymbol{K}(\boldsymbol{x_i},\boldsymbol{x}) \quad (11)$$

4. 実験

本稿で提案した手法の判別分類精度の評価を行うために, UCI Machine Learning Repository (http://www.ics.uci.edu/~mlearn/MLRepository.html) で公開されている Promoter データセットを用いて評価実験を行った.

Promoter データセットは, 塩基を表す $\{A,C,G,T\}$ からなる長さ 57 の文字列データであり, Promoter を含む場合は「+」クラス, 含まない場合は「−」クラスとしたもので, 事例数は各クラス 53 個で計 106 個となっている. 今回の実験では, 各クラスのデータをに 4 分割 (それぞれ A,B,C,D と呼ぶ) し, 75%を訓練データ, 25%を評価データとして計 16 パターンで評価を行った.

その実験結果 (CK*) の一部を編集距離を類似度に利用した研究結果 (EK)[3] と共に表 1 に示す.「縮約無」は訓練データの縮約処理を行わない場合の分類誤り率であり,「縮約有」は訓練データの縮約処理を行った場合の分類誤り率を示している. なお, 評価データに関しては圧縮処理は行っていない.

表 1: EK と CK*における分類誤り率 (%) の比較

訓練 [+ −]	評価 [+ −]	縮約無 EK	縮約有 EK	縮約無 CK*	縮約有 CK*
ABC - ABC	D - D	10.71	17.86	10.71	0.00
ABC - ABD	D - C	10.99	14.56	11.26	10.71
ABC - ACD	D - B	7.14	7.14	7.14	3.57
ABC - BCD	D - A	14.84	17.86	11.26	7.14
ABD - ABC	C - D	7.14	15.11	10.71	0.00
ABD - ABD	C - C	11.86	11.54	7.69	15.38
ABD - ACD	C - B	3.85	11.54	0.00	3.85
ABD - BCD	C - A	11.54	11.54	3.85	11.54
ACD - ABC	B - D	14.29	3.85	7.14	0.00
ACD - ABD	B - C	19.23	15.38	7.69	11.54
ACD - ACD	B - B	7.69	3.85	0.00	0.00
ACD - BCD	B - A	15.38	7.69	7.69	7.69
BCD - ABC	A - D	11.54	18.41	7.14	3.85
BCD - ABD	A - C	11.54	19.23	15.38	19.23
BCD - ACD	A - B	7.69	11.54	3.85	3.85
BCD - BCD	A - A	11.54	15.38	7.69	3.85
平均		11.06	12.66	7.45	6.39

EK の評価に用いた縮約データの縮約率として, 配列数及びデータ量をそれぞれ平均 66%, 51%とした場合, 平均分類誤り率は縮約しない場合とした場合で, それぞれ 11.06%(分散 12.79), 12.66%(分散 22.18) となっている. また, CK の評価に用いた縮約データの縮約率として, 配列数及びデータ量をそれぞれ平均 62%, 57%とした場合, 平均分類誤り率は縮約しない場合とした場合で, それぞれ 7.45%(分散 14.74), 6.39%(分散 30.40) となった. これは, この程度の縮約であれば平均的には, ほぼ類似の性能が得られているが, 縮約データを用いることは, 評価データによるばらつきが増えることを示している.

5. 結論

縮約構造と CK の拡張に基づくカーネル関数による分類手法を提案した. 本稿で提案した手法の適用例から, 冗長構造を除去することで, 縮約構造を利用しない場合より性能向上の可能性のあることが示唆された. 今後の課題として, 縮約構造の獲得方法, カーネル関数の定義, 他の判別器の検討などを通して, 本手法の厳密な検討と確立を図っていきたいと考えている.

参考文献

[1] K. Tsuda, T. Kin, and K. Asai: Marginalized kernels for biological sequences, Bioinformatics, Vol.18, pp.S268-S275, 2002.

[2] S. Mika, G. Rätsh, J. Weston, B. Schölkopf, and K.-R. Müller: Fisher discriminant analysis with kernels, *Neural Networks for Signal Processing IX*, pp.41-48, IEEE, 1999.

[3] 大野 博之, 稲積 宏誠 "縮約構造を利用した分類手法の検討", 第 20 回人工知能学会全国大会, 3A3-2, 人工知能学会,2006.