3A3-02

縮約構造を利用した分類手法の検討

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

大野 博之[*1] 稲積 宏誠[*1]
Hiroyuki Oono Hiroshige Inazumi

[*1]青山学院大学 理工学部 情報テクノロジー学科
College of School of Science and Engineering, Aoyama Gakuin University

We present a new classification method with compressed structure of sample data, where the classifier is constructed using kernel function based on edit distance between compressed structure of learning data and raw sample data. Some experimental results suggest the applicability and effectiveness of our proposed method.

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，より少ない情報量で未知の塩基配列の分類を行う手法を検討する．特に，塩基配列の類似度の尺度として利用可能な編集距離を縮約構造を用いて定義し，分類問題で有効視されているカーネル関数を用いた機械学習による判別分類を実現し，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 [Müller 01]，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

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

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．

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

連絡先: 青山学院大学 理工学部 情報テクノロジー学科
〒 229-8558 神奈川県相模原市淵野辺 5-10-1
E-Mail: oono@ina-lab.it.aoyama.ac.jp

$\boldsymbol{\phi}$ 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．なお，K は以下の条件を満たす必要がある．

- (対称性) $K(\boldsymbol{x}, \boldsymbol{x}')$ が $\boldsymbol{X}$ 上で対称関数となる
- (半正定値性) 任意の $\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n \in \boldsymbol{X}$ に対して，行列 $\boldsymbol{K} = (K(\boldsymbol{x}_i, \boldsymbol{x}_j))_{i,j=1}^{n}$ が半正定値となる

この条件を満たすカーネル関数としては，多項式カーネル (polynomial kernel) や RBF カーネル (Radial Basis Function kernel) がよく知られている．

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. 抽出した縮約構造と塩基配列との類似度を編集距離に基いて定義し，カーネル関数を作成する．
3. 訓練データを用いて，カーネル関数による判別器を作成し，検査データに対して分類・評価する．

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. 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, それ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抽出する. クラスタリングには, 系統樹推定解析を利用し, ボトムアップ法を用いて適当なしきい値によってクラスタ分けをした. また, アラインメントの処理には ClustalW を利用した.

3.2 編集距離を利用したカーネル関数と判別分類

配列の類似性を文字列マッチングの問題として捉えるならば, その類似度は編集距離 (edit distance) [Duda 01] で評価できると考えられる. 編集距離は文字列 $\boldsymbol{x}$ を $\boldsymbol{y}$ へ変換するために必要な基本操作の回数で表され, 基本操作は次の 3 つで与えられる.

- 削除: $\boldsymbol{x}$ 中の文字を削除し, $\boldsymbol{x}$ の長さが 1 文字だけ減る
- 挿入: $\boldsymbol{y}$ 中の文字を $\boldsymbol{x}$ に挿入し, $\boldsymbol{x}$ の長さが 1 文字増える
- 置換: $\boldsymbol{x}$ 中の文字を対応する $\boldsymbol{y}$ 中の文字で置換する

このとき, 2 配列間の編集距離 $D(\boldsymbol{x}[1..m], \boldsymbol{y}[1..n])$ は以下のように定式化できる.

$$D(\boldsymbol{x}[1..m], \boldsymbol{y}[1..n]) = \min \begin{cases} D(\boldsymbol{x}[1..m-1], \boldsymbol{y}[1..n]) + del \\ D(\boldsymbol{x}[1..m], \boldsymbol{y}[1..n-1]) + ins \\ D(\boldsymbol{x}[1..m-1], \boldsymbol{y}[1..n-1]) + \\ \qquad p(x[m], y[n]) \end{cases} \tag{6}$$

ここで, del, ins, p はそれぞれ削除, 挿入, 置換のコストであるが, 基本操作の性質上, それぞれのコストが異なると, コストの低い基本操作が一方的に選択されてしまうため, すべて同じコスト値を持つものとする. しかし, 削除・挿入は比較する配列同士の配列長に影響されるため, 長さに関係なく局所的な類似性の組み合わせを評価するためには, 削除及び挿入のコストよりも置換のコストを大きくする必要がある. そこで, 求められた距離 D には置換操作の回数 ($count\_replace$) をコストとして加え, D は次式 (7) のように修正することとした.

$$D = D + count\_replace \times \alpha \tag{7}$$

この編集距離 D を用いて, n 個の訓練データ $\{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\} \in X$ と, その $n'(n \geq n')$ 個の縮約構造 $\{\boldsymbol{x}'_1, \boldsymbol{x}'_2, \cdots, \boldsymbol{x}'_{n'}\} \in X'$ に対して, 特徴ベクトル $\boldsymbol{\phi}(\boldsymbol{x}) = \{h_{\boldsymbol{x}\boldsymbol{x}_1}, h_{\boldsymbol{x}\boldsymbol{x}_2}, \cdots, h_{\boldsymbol{x}\boldsymbol{x}_n}\}$ の i 番目の要素 $h_{\boldsymbol{x}\boldsymbol{x}_i}$ を

$$h_{\boldsymbol{x}\boldsymbol{x}_i} = e^{-\lambda D(\boldsymbol{x}, \boldsymbol{x}'_i)} \tag{8}$$

と表すことができる. これにより, 編集距離から類似度への変換を行い, さらにパラメータ λ によってスケールの調節が行われる. このとき, 2 つの配列 $\boldsymbol{y}, \boldsymbol{z}$ の類似度としてのカーネル関数 K_{ed} を特徴ベクトルの内積で定義する.

$$K_{ed}(\boldsymbol{y}, \boldsymbol{z}) = (\boldsymbol{\phi}(\boldsymbol{y}) \cdot \boldsymbol{\phi}(\boldsymbol{z})) \tag{9}$$

判別分類は, 上記のカーネル関数を Kernel Fisher Discriminant Analysis(KFDA) [Mika 99] に適用して行い, 判別関数に対する閾値は, 訓練データを判別関数にかけて得られる値の平均値を算出し, その中間値とする.

4. 実験

本稿で提案した手法の判別分類に対する評価を行うために, UCI レポジトリ (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, クラスは Promoter を含んでいれば「+」, 含んでいなければ「−」の 2 つである. データの事例数は各クラス 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 {A,B,C,D} に 4 分割し, 75%を訓練データ, 25%を検査データとして全計 16 パターンで評価を行った.

その実験結果の一部を表 1 に示す.「縮約なし」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,「縮約あり」は訓練した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.

表 1: 分類誤り率による分類性能の比較 (一部抜粋)

訓練 [+ −]	テスト [+ −]	縮約なし (%)	縮約あり (%)
ABC - ABD	D - C	10.99	14.56
ABC - ACD	D - B	7.14	7.14
ABC - BCD	D - A	14.84	17.86
ABD - ABD	C - C	11.86	11.54
ABD - BCD	C - A	11.54	11.54
ACD - ABC	B - D	14.29	3.85
ACD - ABD	B - C	19.23	15.38
ACD - ACD	B - B	7.69	3.85
ACD - BCD	B - A	15.38	7.69

縮約データとしては, 配列数を 66%, データ量を 51%とした場合, 縮約構造を使用しない場合の平均分類誤り率は 11.06%(分散 12.79), 縮約構造を利用した場合の平均分類誤り率は 12.66%(分散 22.18) となった. これは, この程度の縮約では平均的にはほぼ類似の性能が得られているが, 検査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

縮約構造と編集距離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, より少ない情報量ながら, 縮約構造を利用しない場合とほぼ同等の性能が実現でき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

[Müller 01] Klaus-Robert Müller, Sebastian Mika, Gunnar Rätsch, Koji Tsuda, and Bernhard Schölkopf: An Introduction to Kernel-Based Learning Algorithm, *IEEE Transactions on Neural Networks* Vol.12 No.2 pp.181-201, IEEE, 2001.

[Duda 01] Richard O. Duda, Peter E. Hart, and David G. Stork: Pattern Classification 2nd edition, *Wiley-Interscience*, 2001.

[Mika 99] 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.