Vol.2014-MPS-98 No.13
Vol.2014-BIO-38 No.13
2014/6/26

文字列の距離空間上の最大マージン識別器とそのタンパク質科学への応用

小谷野 仁[1,a)]　林田 守広[2,b)]　阿久津 達也[2,c)]

概要：これまでデータと言えば，数や数ベクトルが大部分を占めていたが，近年，計算機科学や生物学において，テキストデータや生物配列など，大量の文字列データが生成されるようになり，文字列データの分類問題は，様々な領域に共通の問題となっている．この問題に対して現在最もよく用いられている方法は，文字列カーネルによって文字列を数ベクトルに変換し，それにサポートベクターマシーンを適用することである．しかし，この変換は 1 対 1 ではなく，文字列を構成する文字の並びに関するかなりの量の情報を捨ててしまう．また，この接近法のより重要な問題は，学習機械を訓練し，テストするために与えられたデータはある確率法則に従って生成された文字列であるという重要な側面を考慮し，確率論を用いて学習機械の汎化誤差を理論的に評価することを不可能にしていることである．なぜ，文字列データを分類するために，それを数ベクトルに変換し，数ベクトル空間上で動作する学習機械を用いるのだろうか．文字列を分類するには，文字列の集合上で動作する学習機械を用いるのが自然だろう．我々は，文字列を数ベクトルに変換せずに，文字列自体を入力として受け取る学習機械を構築することにより，この分類問題に接近した．このような学習機械の汎化誤差を理論的に評価するには，文字列に対する確率論が必要である．文字列は, これまで, 数学の対象というよりは, 計算機科学の対象であり, 文字列の集合に位相構造や代数構造を与えて, その上で確率論を展開するということはなされてこなかったが, 著者等のうちの 1 人と彼の共同研究者は, 以前の研究において, Levenshtein 距離が与えられた文字列の距離空間上で確率論を展開して, ベクトル空間における大数強法則の, この空間におけるアナロジーを証明した．この研究において, 我々は, この文字列の集合上の確率論を応用することにより, ある正則条件の下で, 我々の学習機械が漸近的に最適な仕方で文字列を分類することを証明した．更に，我々の学習機械を，アミノ酸配列を用いたタンパク質間相互作用の予測問題に応用して，実際のデータ解析におけるその有用性を示した．

キーワード：文字列の分類, 機械学習, 文字列の距離空間上の確率論, タンパク質間相互作用の予測

Maximum Margin Classifier Working in a Metric Space of Strings and Its Application to Protein Science

Hitoshi Koyano[1,a)]　Morihiro Hayashida[2,b)]　Tatsuya Akutsu[2,c)]

Abstract: Numbers and numerical vectors account for a large portion of data. However, recently, the amount of string data generated has increased dramatically. Consequently, classifying string data is a common problem in many fields. The most widely used approach to this problem is to convert strings into numerical vectors using string kernels and subsequently apply a support vector machine that works in a numerical vector space. However, this non-one-to-one conversion involves information loss and makes it impossible to evaluate, using probability theory, the generalization error of a learning machine, considering that the given data to train and test the machine are strings generated according to probability laws. We approach this classification problem by constructing a classifier that receives the strings themselves as inputs. To evaluate the generalization error of such a classifier theoretically, probability theory for strings is required. A string is an object of computer science rather than mathematics, and probability theory for strings has not been constructed. However, one of the authors and his colleague, in previous studies, first developed a probability theory on a metric space of strings provided with the Levenshtein distance and demonstrated an analogy of the strong law of large numbers in a numerical vector space. In this study, by applying this probability theory on a set of strings, we demonstrate that our developed learning machine classifies strings in an asymptotically optimal manner. Furthermore, we demonstrate the usefulness of our machine in practical data analysis by applying it to predicting protein-protein interactions using amino acid sequences.

Keywords: Classifying strings, machine learning, probability theory on a metric space of strings, predicting protein–protein interactions

Vol.2014-MPS-98 No.13
Vol.2014-BIO-38 No.13
2014/6/26

1. はじめに

これまでデータと言えば, 数と数ベクトルが大部分を占めていた. しかし, 近年, Web 上に大量のテキストデータが生産されている. また, 生物学の領域では, 遺伝子, RNA 及びタンパク質に関する大量のデータが生成されているが, これらはヌクレオチドやアミノ酸の配列であり, 文字列として表される. この結果, 文字列の分類問題は, 計算機科学や生物学を含む様々な領域に共通の問題となっている. この問題に対して現在最もよく用いられている方法は, 文字列カーネルにより文字列を数ベクトルに変換し, それにサポートベクターマシーン (SVM) (例えば [1], [4], [7], [9], [35] を参照) を適用することである. 初期の文字列カーネルは [12], [27], [38] によって開発されたが, これらの論文は, 文字列の間の類似度をそれらに共通の部分列の数に基づいて定義するという考えを打ち出している. [23], [30] は, ギャップが入ることが許される共通の部分列は考慮せずに, 共通の部分文字列の数に基づいて, 文字列の間の類似度を数値化する文字列カーネルであるスペクトルカーネルを用いている. スペクトルカーネルは, その後, [21], [22], [37] によって拡張された. これらのカーネルの他にも, [26], [33], [36], [40] などにより多くの文字列カーネルが開発され, バイオインフォマティクスの問題に応用されている. 文字列を構成する文字の並びに関するかなりの量の情報を捨ててしまうが, 様々な文字列カーネルの中で最も広く使われているのは, スペクトルカーネルだろう. 一般に, 文字列を数ベクトルに変換することは情報の損失を伴う. それにも関わらず, 文字列を分類するために, なぜ文字列を数ベクトルに変換し, 数ベクトル空間上で動作する識別器を用いるのだろうか. 文字列を分類するには, 文字列の集合上で動作する識別器を用いるのが自然だろう. しかし, SVM の入力は基本的に数ベクトルであるため, これまでに, 学習機械に文字列自体を入力として与えることはあまり検討されてこなかった. また, 従来の接近法のより重要な問題は, それが, 学習機械の訓練とテストのために与えられたデータはある確率法則に従って生成された文字列であるという重要な側面を無視していることである. この結果, ある特定の数値実験において, またはある特定のデータセットに適用した際に, 他の学習機械より良い結果を出したかどうかによって, 学習機械の性能が評価され, その汎化誤差を理論的に評価するという学習機械の最も基本的な仕方での評価は, 事実上放棄されてきた.

この研究において, 我々は, 文字列の距離空間上で動作する SVM のアナロジーを構築することにより, 文字列を数ベクトルに変換することなく, 文字列を分類するための方法を開発した. 文字列の集合上で動作する識別器の汎化誤差を理論的に評価するためには, 文字列に対する確率論が必要である. 数学は, その長い歴史の中で, 数, 多様体, 演算, 方程式, 関数, 作用素など, 多くの対象について深い研究をしてきたが, 文字列はほとんど研究してこなかった. 文字列は, 数学の対象というよりは, 計算機科学が深く研究してきた対象である. 計算機科学の 1 領域である stringology は, 文字列処理のためのアルゴリズムとデータ構造について徹底的に研究してきた (例えば [8], [11] を参照). しかし, 計算機科学は, ある対象の集合を考え, それに距離や演算を定義することにより位相構造や代数構造を与え, その集合の上の関数, 作用素, 確率などを研究するという, 数学がとっている研究方法では, 文字列を研究してこなかった. [18] は, Levenshtein 距離 [24] が与えられた文字列の距離空間上で初めて確率論を展開し, 確率文字列の列に対する大数強法則のアナロジーと確率文字列に対する分散の漸近挙動に関する結果を証明した. これらの結果は, 文字列データに対する統計理論を開発するための基本的な道具を与える. [18] は, この確率論に基づいて統計的方法を開発し, それを生物配列の解析に応用している. この研究において, 我々は, 文字列の距離空間上に作られたこの確率論を応用することにより, 我々が開発した学習機械の汎化誤差の理論的な評価を与える. 更に, 我々の学習機械をアミノ酸配列を用いたタンパク質間相互作用の予測問題に応用して, 実際のデータ解析におけるその有用性を示す.

2. 問題の定式化

SVM は, (i) ベクトル空間の双対空間上での学習, (ii) 入力ベクトルからの特徴の抽出, 及び (iii) カーネル関数の使用など, いくつかの特徴を持っているが, 以下では, 我々は, マージン最大化原理の下で超平面を選ぶことにより, 空間を 2 つの排反な部分集合の直和に分割する識別器を SVM と呼ぶ. $\mathbb{R}$ と $\mathbb{R}^p$ をそれぞれ実数の集合と p 次元実ベクトル空間とする. 簡単のため, 平面 $\mathbb{R}^2$ を考える. $\mathbb{R}^2$ における直線は, $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^2$ に対して $\{\boldsymbol{z} \in \mathbb{R}^2 : \boldsymbol{z} = \alpha\boldsymbol{x} + \beta\boldsymbol{y}, \alpha + \beta = 1, \alpha, \beta \in \mathbb{R}\}$ 及び $a, b \in \mathbb{R}$ に対して $\{(x, y) \in \mathbb{R}^2 : y = ax + b, x \in \mathbb{R}\}$ と表される. 1 番目の表現においては加法とスカラー乗法が使われているから, それは $\mathbb{R}^2$ のベクトル空間としての構造を使って, 直線を表現している. 一方で, 2 番目の表現においては加法と乗法が使われているから, それは $\mathbb{R}$ の体としての構造を使って, 直線を表現している. アルファベット $A = \{a_1, \cdots, a_{c-1}\}$ 上の文字列の集合を A^* によって表す.

[1] 京都大学大学院医学研究科臨床研究総合センター
Institute for Advancement of Clinical and Translational Science, Graduate School of Medicine, Kyoto University, 54 Kawahara-cho, Shogoin, Sakyo-ku, Kyoto 606-8507, Japan
[2] 京都大学化学研究所バイオインフォマティクスセンター
Bioinformatics Center, Institute for Chemical Research, Kyoto University, Gokasho, Uji, Kyoto 611-0011, Japan
a) koyano@kuhp.kyoto-u.ac.jp
b) morihiro@kuicr.kyoto-u.ac.jp
c) takutsu@kuicr.kyoto-u.ac.jp

A^* 上の内在的な算法と距離はそれぞれ連接 (以下 $\cdot$ によって表す) と Levenshtein 距離 (以下 d_L によって表す) だろう. そこで, $\cdot$ と d_L によって A^* に代数構造と位相構造を与える. そうすると, A^* は非可換位相半群をなすが, ベクトル空間や体にはならないから, 上の 2 つの形式を用いて A^* に “直線” を定義することはできない. しかし, これは, A^* には直線を定義できないということは意味しない. そこで, 次の 2 つの問題を考えよう: (i) 何らかの方法で A^* に “直線” を定義できるだろうか. (ii) それができたとして, その “直線” を使って, A^* を 2 つの背反な部分集合の直和に分割できるだろうか. 1 番目の問題に対する答えは “Yes” であり, 2 番目の問題に対する答えは “No” である.

空間の曲線とは, その空間のある点をそれに隣接する点のうちの 1 つに結合するという操作を繰り返すことによって得られる, その空間の部分集合のことと見なすと, A^* の “曲線” を, 例えば次のように定義することができる: $s_1, \cdots, s_n \in A^*$ に対して $d_L(s_i, s_{i+1}) = 1$, $i = 1, \cdots, n-1$ が成り立つならば, $\{s_1, \cdots, s_n\}$ を A^* における “曲線” と呼ぶ. 更に, 空間の 2 つの点の間の線分とは, それらの 2 つの点を結ぶ最短の曲線のことと見なすと, A^* における “線分” を, 例えば次のように定義することができる: $s, s' \in A^*$ に対して $d_L(s, s') = n$ とする. そうすると, 挿入, 削除, 及び置換の 3 種類の操作を n 回行うことによって, s は s' に変形され得る. 各 $i = 1, \cdots, n-1$ に対して, s にこの n 回の操作のうちの最初の i 回を行うことによって得られる文字列を $s_{(i)}$ によって示す. 文字列を結ぶ “線分” の一意性のため, 挿入, 削除, 及び置換の間に優先順位が与えられているとし, s に対する一連の操作は, s を構成する文字の順序に従って，また 3 種類の操作の優先順位に従って行われるという約束をおく. この時, $\{s, s_{(1)}, \cdots, s_{(n-1)}, s'\}$ を A^* における s と s' を結ぶ “線分” と呼ぶ.

そこで, A^* 全体ではないが, 十分に長い文字列 s を選び, 空文字列 (0 個の文字からなる文字列) と s を結ぶ線分を引くことにより, s の長さ以下の長さを持つ文字列からなる, 応用上十分に大きな A^* の部分集合を分割することを考えよう. アルファベット $A = \{a_1, \cdots, a_{c-1}\}$ は Hamming 距離 $d_H(a_i, a_j) = 0$ ($i = j$ の時) または 1 ($i \neq j$ の時) によって距離空間になる. これに対して, 実数の集合 $\mathbb{R}$ は差の絶対値 $d(x, y) = |x - y|$ によって距離空間になるが, 同時に通常の “以下である” という関係 $\leq$ によって全順序集合をなす. $\mathbb{R}$ 上の距離 d と全順序 $\leq$ は, $x, y, z \in \mathbb{R}$ に対して, $x \leq y$ かつ $y \leq z$ ならば, $d(x, y) \leq d(x, z)$ が成り立つという意味で整合的である. A 上には, $\mathbb{R}$ 上の大小関係 $\leq$ のような内在的な全順序は存在しない. それでは, 上で述べた意味で Hamming 距離 d_H と整合的な全順序を A 上に定義することにより, A が既に持っている距離空間としての構造を破壊せずに, A を全順序集合にすることはできないだろうか. d_H の定義から, これは明らかに不可能である. この結果, 次の問題が生じる. $\mathbb{R}^2$ は, 直線 $\ell = \{(x, y) \in \mathbb{R}^2 : y = ax + b\}$ によって, 上半空間 $H^+ = \{(x, y) \in \mathbb{R}^2 : ax + b \leq y\} - \ell$ と下半空間 $H^- = \{(x, y) \in \mathbb{R}^2 : y \leq ax + b\} - \ell$ に分割されるが, H^+ と H^- は, $\mathbb{R}$ 上の全順序 $\leq$ を使って定義されている. つまり, 直積空間 $\mathbb{R}^2$ における直線の上側と下側の領域という概念が意味を持つには, 直積因子 $\mathbb{R}$ 上の全順序が必要である. Levenshtein 距離を使って, 上述の仕方で A^* に曲線と線分のアナロジーを定義することができた. しかし, A 上に Hamming 距離と整合的な全順序を定義できないため, A の距離空間としての構造を破壊せずに, A における線分の上側と下側という概念に意味を持たせることができない. この結果, $\mathbb{R}^2$ においてと異なり, 直線のような “閉じていない” 部分集合を決定することによっては A^* の直和分割は作れない.

しかし, 上の議論は, どのようにしても A^* を 2 つの排反な部分集合に分割することはできないということは意味しない. 位相幾何学における Jordan の曲線定理 [17] や Jordan–Brouwer の分離定理 [5] が述べているように, $\mathbb{R}^2$ と $\mathbb{R}^p$ ($p \geq 3$) は, それぞれ直線と超平面を決定せずとも, 閉曲線と超球面を選ぶことにより, 2 つの排反な部分集合に分割される. 直線を引く以外の方法で, A^* を 2 つの排反な部分集合に分割することはできないだろうか. $s \in A^*$ と $r \in \mathbb{Z}^+$ に対して $U(s, r) = \{t \in A^* : d_L(t, s) \leq r\}$ と定め ($\mathbb{Z}^+$ は正の整数の集合), A^* を $U(s, r)$ と $U(s, r)^c = A^* - U(s, r)$ に分割することを考えよう. つまり, A^* に球面を描き, A^* をその内部と外部に分割するのである. この仕方の分割は, 下側や上側という概念を必要としない. 以下では, 我々は, $\partial U(s, r) = \{t \in A^* : d_L(t, s) = r\}$ を識別球面と呼び, $U(s, r)$ に含まれる文字列の数を $\partial U(s, r)$ の大きさと言う.

3. 文字列の距離空間上で動作する SVM のアナロジー

前節で述べた仕方で A^* を分割するには, 正例と負例が与えられた時, 識別球面 $\partial U(s, r)$ の中心 $s \in A^*$ と半径 $r \in \mathbb{Z}^+$ を指定する必要がある. 正例 $X_m = \{s_1, \cdots, s_m\}$ と負例 $Y_n = \{t_1, \cdots, t_n\}$ は, $s_0 \in A^*$ が存在して, $\max_{1 \leq i \leq m}\{d_L(s_i, s_0)\} < \min_{1 \leq i \leq n}\{d_L(t_i, s_0)\}$ が成り立つ時, 球形分離可能であると言い, 球形分離可能でない時, 球形分離不可能であると言うことにする. コンセンサス配列が一意に定まる m 個の文字列の組の集合を $[(A^*)^m]$ によって表す (コンセンサス配列の形式的な定義については, [18] を参照). 以下では, $s_1, \cdots, s_m \in [(A^*)^m]$ と仮定し, 識別球面の中心としては, 正例 $s_1, \cdots, s_m$ のコンセンサス配列 $\bar{s}_m$ を選ぶ.

まず, 正例と負例が球形分離可能である場合に, 識別球面の半径を選ぶ問題を考える. $\mathbb{R}^p$ 上の SVM の識別超平面に対して同様に, ある文字列と識別球面の間の距離は,

その文字列とそれに最も近い球面上の文字列の間の距離のことである. また, 正例と負例が与えられている時, 識別球面のマージンとは, その球面とそれに最も近い例の間の距離のことである. マージン最大化原理の下では, 次の結果が直ちに得られる: 正例 $X_m = \{s_1, \cdots, s_m\}$ と負例 $Y_n = \{t_1, \cdots, t_n\}$ が $\bar{s}_m$ に関して球形分離可能であるならば, マージンを最大化する識別球面の半径は,

$$r^* = \frac{1}{2}\Big\{ \max_{1\le i\le m} \{d_L(s_i, \bar{s}_m)\} + \min_{1\le i\le n} \{d_L(t_i, \bar{s}_m)\}\Big\} \quad (1)$$

によって与えられる. r^* が整数でないならば, r^* に最も近い整数のうちの 1 つを任意に選ぶ.

次に, 正例 X_m と負例 Y_n が球形分離不可能である場合を考える. 中心 $\bar{s}_m$ と半径 r を持つ識別球面 $\partial U(\bar{s}_m, r)$ が正しく識別する正例の部分標本と負例の部分標本を, それぞれ $X_m(\bar{s}_m, r)$ と $Y_n(\bar{s}_m, r)$ によって表す. $\sharp S$ を有限集合 S の元の数とする. X_m と Y_n に属する文字列で, $\partial U(\bar{s}_m, r)$ が誤って識別するものの数は, それぞれ $m - \sharp X_m(\bar{s}_m, r)$ と $n - \sharp Y_n(\bar{s}_m, r)$ と表せる. 2 つの標本 X_m と Y_n が球形分離不可能である時には, ソフトマージン最適化の場合に $\mathbb{R}^p$ 上の通常の SVM によって用いられる原理を若干修正して, 誤識別数最小化とマージン最大化を組み合わせた次の原理に基づいて, 識別球面の半径を選ぶことにする. 正例と負例が球形分離可能であるならば, 次の手続きは, 式 (1) に従って半径を選ぶことと同じである.

ステップ 1 (誤識別数最小化). 誤って識別される入力の数を最小化する半径の集合を求める. すなわち, $\tilde{r} = \arg\min_{r\in\mathbb{Z}^+}\{m - \sharp X_m(\bar{s}_m, r) + n - \sharp Y_n(\bar{s}_m, r)\}$, あるいは同じことであるが, $\tilde{r} = \arg\max_{r\in\mathbb{Z}^+}\{\sharp X_m(\bar{s}_m, r) + \sharp Y_n(\bar{s}_m, r)\}$ を満たす正の整数 $\tilde{r}$ の集合を求める. この集合を $\tilde{R}$ によって表す. $\tilde{R}$ は非空な有限集合である.

ステップ 2 (マージン最大化). 正しく識別される文字列のうち最も近いものまでの距離を最大化する $r^* \in \tilde{R}$ を選ぶ (そのような r^* が一意に定まらない場合には, それらのうちの 1 つを任意に選ぶ). このステップは形式的には次のように書ける: $s \in X_m(\bar{s}_m, r)$ と $\partial U(\bar{s}_m, r)$ の間の距離と $t \in Y_n(\bar{s}_m, r)$ と $\partial U(\bar{s}_m, r)$ の間の距離は, それぞれ $r - d_L(s, \bar{s}_m)$ と $d_L(t, \bar{s}_m) - r$ と等しい. $\mathbb{R}^p$ 上の通常の SVM に対するサポートベクトルに対してと異なり, s と t がサポート文字列である時, これらの距離の和 $r - d_L(s, \bar{s}_m) + d_L(t, \bar{s}_m) - r = d_L(t, \bar{s}_m) + d_L(s, \bar{s}_m)$ は奇数である場合があるから, これらの距離は必ずしも等しくないことに注意する. 最適は半径 r^* は,

$$\rho(\tilde{r}) = \min_{(s,t)\in X_m(\bar{s}_m,\tilde{r})\times Y_n(\bar{s}_m,\tilde{r})} \min\{\tilde{r} - d_L(s, \bar{s}_m), d_L(t, \bar{s}_m) - \tilde{r}\},\ \tilde{r} \in \tilde{R}$$

に対して $r^* = \arg\max_{\tilde{r}\in\tilde{R}} \rho(\tilde{r})$ と表せる. また, サポート文字列は

$$(s^*, t^*) = \arg \min_{(s,t)\in X_m(\bar{s}_m,r^*)\times Y_n(\bar{s}_m,r^*)} \min\{r^* - d_L(s, \bar{s}_m), d_L(t, \bar{s}_m) - r^*\}$$

と表せる. 集合 $\tilde{R}$ を求めるには, $\max_{1\le i\le m}\{d_L(s_i, \bar{s}_m)\}$ と $\min_{1\le i\le n}\{t_i, d_L(\bar{s}_m)\} - 1$ の間の半径のみを調べれば十分である.

4. 提案された学習機械の漸近的最適性

前節で構築された SVM の A^* 上のアナロジーは, 汎化誤差の観点から何らかの最適性を持つだろうか. 本節において, 我々は, [18] によって提案された A^* 上の確率論の基本的な枠組みとその枠組みの中で証明されたコンセンサス配列の漸近的な挙動に関する結果を応用することにより, この問題を考察する. 本節で使われる定義と定理については, [18] を参照して欲しい. $\boldsymbol{p}_1$ と $\boldsymbol{p}_2$ をそれぞれ正例と負例を生成する分布の確率関数とする. $\boldsymbol{p}_1$ のコンセンサス配列を $\boldsymbol{m}'_1$ によって表す. D_1 と D_2 をそれぞれ $\boldsymbol{p}_1$ と $\boldsymbol{p}_2$ の台とする. すなわち, $i = 1, 2$ に対して $D_i = \{s \in A^* : \boldsymbol{p}_i(s) > 0\}$. $D_1 - D_2 \neq \emptyset$ かつ $D_2 - D_1 \neq \emptyset$ と仮定する. $D_1 \cap D_2 = \emptyset$ ならば, 有限回の学習の後, 汎化誤差が 0 になる確率は 1 である. 従って, 以下では $D_1 \cap D_2 \neq \emptyset$ の場合を考える. 識別球面 $\partial U(s, r)$ の汎化誤差 $E_0(s, r)$ は,

$$E_1(s,r) = \sum_{t\in D_1\cap U(s,r)^c} \boldsymbol{p}_1(t),\ E_2(s,r) = \sum_{t\in D_2\cap U(s,r)} \boldsymbol{p}_2(t). \quad (2)$$

に対して $E_0(s, r) = E_1(s, r) + E_2(s, r)$ と書ける. 各 $s_0 \in A^*$ に対して, 形式的に

$$(s^\dagger, r^\dagger) = \arg \min_{(s,r)\in A^*\times\mathbb{Z}^+} E_0(s, r),$$
$$r^\dagger(s_0) = \arg \min_{r\in\mathbb{Z}^+} E_0(s_0, r)$$

と定める. $r^\dagger(s_0)$ は, 中心が与えられている時に, 汎化誤差の観点から最適な識別球面の半径である. 任意の $t \in A^*$ に対して, X_m と Y_n における t の相対頻度をそれぞれ $\hat{\boldsymbol{p}}_1(t)$ と $\hat{\boldsymbol{p}}_2(t)$ によって表す. $s \in A^*$ と $r \in \mathbb{Z}^+$ に対して

$$\hat{E}_1(s,r) = \sum_{t\in X_m\cap U(s,r)^c} \hat{\boldsymbol{p}}_1(t),\ \hat{E}_2(s,r) = \sum_{t\in Y_n\cap U(s,r)} \hat{\boldsymbol{p}}_2(t) \quad (3)$$

と定める.

まず, 識別球面の中心としては $\bar{s}_m$ が使われると仮定し, 我々の学習機械が学習の過程で r^* を更新するに従って, r^* が汎化誤差の観点から最適な半径に収束するかという問題を考える. [18] の定理 2 の条件の下では, 十分に多くの正例が与えられている時, $\bar{s}_m$ は確率 1 で $\boldsymbol{m}'_1$ と等しいから, 最適な半径は $r^\dagger(\boldsymbol{m}'_1)$ であることに注意する.

定理 1 (r^* の漸近的最適性) (i) 正例 $s_1, \cdots, s_m$ が独立で同一の確率関数 $\boldsymbol{p}_1$ を持つ確率文字列 $\sigma_1, \cdots, \sigma_m$ の実現値であって, 負例 $t_1, \cdots, t_n$ が独立で同一の確率関数 $\boldsymbol{p}_2$

を持つ確率文字列 $\tau_1, \cdots, \tau_n$ の実現値であり, (ii) [18] の定理 2 の条件が満たされており, (iii) $r^\dagger(\boldsymbol{m}'_1)$ が一意に存在するならば,

$$r^* \xrightarrow{\text{a.s.}} r^\dagger(\boldsymbol{m}'_1)\ (m, n \longrightarrow \infty)$$

が成り立つ. ここで, $\xrightarrow{\text{a.s.}}$ は概収束を表している. つまり, 識別球面の中心として $\bar{s}_m$ が与えられている時, r^* は漸近的に最適な半径に確率 1 で収束する.

次に, $\bar{s}_m$ の最適性の問題を考える. 我々は, この問題を, 正例と負例が球形分離不可能である状況をモデル化した次の設定の下で考察する: (i) $\boldsymbol{p}_1(s)$ は, D_1 上で $d_L(s, \boldsymbol{m}'_1)$ に関して単調非増加である. (ii) $d_0 \in \mathbb{Z}^+$ が存在して, $\boldsymbol{p}_2(s)$ は $D'_2 = \{s \in D_2 : d_L(s, \boldsymbol{m}'_1) \leq d_0\}$ 上で $d_L(s, \boldsymbol{m}'_1)$ に関して単調非減少であり, かつ $D_1 \not\supset D'_2$ が成り立つ. d_0 は十分に大きいとし, $(D_1 \cup D'_2)^c$ と交わらない識別球面のみを考察の対象にする. 識別球面の半径としては r^* が選ばれると仮定する. $\sharp U(s, r)$ は, s の長さと r に関して単調に増加することに注意する. 任意の $r \in \mathbb{Z}^+$ に対して, $\sharp U(s', r') = \sharp U(\bar{s}_m, r)$, $\sharp U(s', r') \leq \sharp U(\bar{s}_m, r)$ 及び $\sharp U(s', r') \geq \sharp U(\bar{s}_m, r)$ が成り立つ組 $(s', r') \in A^* \times \mathbb{Z}^+$ の集合を, それぞれ $B_0(r), B_1(r)$ 及び $B_2(r)$ によって表す.

定理 2 ($\bar{s}_m$ の漸近的最適性) 上述の設定の下で, [18] の定理 2 の条件が満たされているならば, 任意の $r \in \mathbb{Z}^+$, $(s', r') \in B_j(r)$ 及び $j \in \{0, 1, 2\}$ に対して, $m, n \longrightarrow \infty$ の時,

$$E_j(\bar{s}_m, r) \leq E_j(s', r') \text{ a.s.}$$

が成り立つ. つまり, 任意の半径 r に対して, 中心 $\bar{s}_m$ を持つ識別球面は, それと等しい大きさを持つ識別球面の族の中で漸近的に最適である. また, それ以下の大きさを持つ識別球面の族の中で漸近的に最小の false negative の確率を持ち, それ以上の大きさを持つ識別球面の族の中で最小 false positive の確率を持つ.

定理 1 と 2 の証明は当日に説明する.

5. タンパク質間相互作用の予測への応用

我々の学習機械は, 十分に大きな訓練標本が与えられている時, 前節で述べられた条件の下では, ほぼ最適な仕方で文字列を分類する. しかし, 文字列の分類問題に対して, 常に大きな訓練標本が得られるとは限らない. そのような場合, 我々の学習機械はどの程度正確に文字列を分類するだろうか. タンパク質は 20 種類のアミノ酸の重合体であり, 20 種類の文字からなるアルファベット上の文字列として表される. 大部分のタンパク質は他のタンパク質と複合体を形成してその機能を発揮するため, タンパク質間相互作用の予測は, バイオインフォマティクスにおける重要な問題のうちの 1 つとなっている. タンパク質のドメインは, 一般に他のタンパク質の複数のドメインと相互作用するが, 他のタンパク質の多くのドメインと相互作用するドメインを持ち, タンパク質間相互作用ネットワークにおいてハブの働きをするタンパク質はごくわずかである [14], [15]. よって, タンパク質間相互作用の予測問題においては, 必ずしも多くの正例が得られない. そこで, 我々は, この予測問題を, タンパク質のドメインを所与のドメインと相互作用するドメインのクラスと相互作用しないドメインのクラスに分類する問題として定式化し, この問題に, 前節で構築した A^* 上の SVM のアナロジーを応用する. そうして, 2-スペクトルカーネルを用いた SVM と比較することにより, 我々の学習機械の分類性能を検討する.

我々は, まず, Protein Data Bank (PDB) [32] から得られた, タンパク質のドメイン間相互作用の 3 次元構造のデータベースである three-dimensional interacting domains (3did) データベース [29] を使って, 正例を作成した. PDB は, X 線結晶構造解析や NMR のような実験から得られたタンパク質とタンパク質複合体の 3 次元構造のデータを含んでいる. 我々は, 10 個のタンパク質のドメインの配列 '1il1 A:134–215' (PDB ID 1il1 のタンパク質の A 鎖中の残基 134 から 215 までのアミノ酸の部分配列), '2bnq E:127–220', '1inq B:1011–1092', '1it9 H:133–216', '1it9 L:123–210', '1ikv A:317–419', '1cff A:84–145', '1iza A:1–20', '1ifh L:119–206' 及び '1p2c F:1501–1627' を選び, これらのそれぞれと相互作用するドメインの配列を 3did から集めた.

次に, 負例を作成する方法を考える. これまで, 文字列データや配列データの識別器の開発と検証においては, 負例として, 正例でなさそうな実際の配列や, ランダムに生成された配列などの人工的な配列が使われてきた. 例えば, *E. coli* のプロモーター領域の予測において, [13] は, 負例としてコード領域からランダムに選んだ配列を使っている. しかし, [20] が示しているように, このような負例の使用は, 現実の生物学的な識別問題と大きく異なっている. また, miRNA と mRNA の相互作用の予測問題において, [10], [16], [39] は, 負例としてランダムに生成された人工的な配列を使っているが, [19], [25], [31] が実験によって示しているように, そのような配列はしばしば miRNA と相互作用し, それらが本当に負例であるかが明らかでない. ランダムに生成された配列が本当に負例であったとしても, それらは正例と非現実的に異なっているかも知れない. この場合, [2] が指摘しているように, 正例と負例が容易に識別され, 他の独立なデータセットに対しては貧弱な識別性能を示す識別器が構築されてしまう可能性がある.

そこで, この研究では, 我々は, 生物物理化学の知見に基づいて, 正例とある程度似ているが, 相互作用しないであろう負例を作成する方法を考えた. [3] は, ヘテロ二量体タンパク質複合体の 2,325 箇所のアラニン変異体のデータベースを作成し, どのアミノ酸が高い頻度でタンパク質複合体

のインターフェイスに位置するかを調べた. また, [28] は, PDB に登録されている 1,629 個の 2 鎖のタンパク質複合体の立体構造のデータに基づいて, 高い頻度で複合体のインターフェイスに位置するアミノ酸を同定した. これらの研究の結果によると, Arg, Asp, Trp 及び Tyr は, 高い頻度でタンパク質複合体のインターフェイスに位置し, 一方で Lys と Glu はあまりインターフェイスに位置しない. Arg と Lys は共に正の電荷を持つが, Arg はインターフェイスに位置する傾向がり, 逆に Lys はインターフェイスに位置する傾向がない. 対照的に, Asp と Glu は共に負の電荷を持つが, Asp はインターフェイスに位置する傾向があるのに対して, Glu はそうでない. これらのことから, 側鎖が比較的低いエントロピーを持つアミノ酸はインターフェイスに位置しやすいと推測される. また, 芳香族アミノ酸の側鎖の π 電子は, 正の電荷を持つアミノ酸と強く相互作用することが知られているが [6], Trp や Tyr のような芳香族アミノ酸が高い頻度でインターフェイスに位置するのは, このためと考えられる.

そこで, 我々は, 次のステップで各正例から負例を作成した: (i) まず, 正例に含まれる Arg を Glu に, Asp, Trp 及び Tyr を Lys に確率 0.5 で置換する (全ての Arg, Asp, Trp 及び Tyr を置換すると, 負例がこれらのアミノ酸を表す 4 種類の文字を含まなくなってしまうため). [34] によると, 一般的なタンパク質における Arg, Asp, Trp 及び Tyr の頻度は, それぞれ 5.1%, 5.3%, 1.4% 及び 3.2% である. 従って, このステップで, 正例を構成する文字の約 7.5% に相当する, タンパク質複合体のインターフェイスに位置する傾向があるアミノ酸を表す文字が, インターフェイスに位置する傾向がないアミノ酸を表す文字に置換される. (ii) 次に, ステップ (i) で置換が行われた正例を 3 等分し, 真ん中の部分文字列の中の 1 つの文字をランダムに選び, 置換が行われた正例の最初の文字から選んだ文字までからなる前半の部分文字列と, 置換が行われた正例の残りの文字からなる後半の部分文字列の順序を入れ替える.

標本の大きさ (正例の数と負例の数の和) は, 表 1 に N として与えられている. 我々は, $N = 20$ の小さな標本が与えられている場合から, $N = 126$ の十分に大きな標本が与えられている場合までを検討した. true positive の数 TP, false positive の数 FP, true negative の数 TN 及び false negative の数 FN に対して, accuracy $= (TP + TN)/N$, precision $= TP/(TP + FP)$, recall $= TP/(TP + FN)$ 及び F-measure $= 2 \times$ precision $\times$ recall$/$(precision + recall) の 4 つの指標を使って, 学習機械の性能を評価した. 正例の標本と負例の標本をそれぞれ等しい大きさの 3 つの部分標本にランダムに分割した後, 2 つの部分標本の合併を訓練標本として用い, 他の部分標本を試験標本として用いて, accuracy, precision, recall 及び F-measure を計算した. そうして, この過程を 50 回繰り返して, これらの各指標の平均を求めた. この手順で得られた 4 つの指標の平均が, 表 1 に示されている. この表から, 我々の学習機械は, 2-スペクトルカーネルを用いた SVM と比較して, 高い識別性能を持っていることが分かる.

6. まとめ

近年, 生成される文字列データの量は劇的に増加している. この結果, ある確率法則に従ってランダムに数を生成する確率変数や, ランダムに関数を生成する確率過程が様々な領域において必須であるのと同様に, ある確率法則に従ってランダムに文字列を生成する確率文字列が必要になってきている. 数データに対する統計学は, 確率論に基づいて厳密に構築されている. これと同様に, テキストマイニングの手法やバイオインフォマティクスにおける生物配列解析の方法に対しても, 文字列の集合上の確率論に基づいた新しい方法の開発や既存の方法の体系化が求められるようになるだろう. 数値実験のみではなく, 理論的な解析にも基づいた方法の開発と理論的な枠組みの下での様々な方法の体系化が, 文字列データの解析に関するこれからの方法論上の研究の重要な課題であるだろう.

参考文献

[1] Aizerman, M. A., Braverman, E. M. and Rozoner, L. I.: Theoretical foundations of the potential function method in pattern recognition learning, *Autom. Remote Control*, Vol. 25, pp. 821–837 (1964).
[2] Bandyopadhyay, S. and Mitra, R.: TargetMiner: microRNA target prediction with systematic identification of tissue-specific negative examples, *Bioinformatics*, Vol. 25, No. 20, pp. 2625–2631 (2009).
[3] Bogan, A. A. and Thorn, K. S.: Anatomy of hot spots in protein interfaces, *J. Mol. Biol.*, Vol. 280, No. 1, pp. 1–9 (1998).
[4] Boser, B. E., Guyon, I. M. and Vapnik, V. N.: A training algorithm for optimal margin classifiers, *Proc. 5th Annu. Workshop Comput. Learn. Theory* (Houssler, D., ed.), pp. 144–152 (1992).
[5] Brouwer, L. E. J.: Beweis des Jordanschen Satzes für den n-dimensionalen Raum, *Mathematische Annalen*, pp. 314–319 (1911).
[6] Burley, S. K. and Petsko, G. A.: Weakly polar interactions in proteins, *Advances in Protein Chemistry* (Anfinsen, C. B., Edsall, J. T., Richards, F. M. and Eisenberg, D. S., eds.), Vol. 39, Academic Press, Waltham, MA, pp. 125–189 (1988).
[7] Cortes, C. and Vapnik, V. N.: Support-vector networks, *Mach. Learn.*, Vol. 20, No. 3, pp. 273–297 (1995).
[8] Crochemore, M. and Rytter, W.: *Jewels of Stringology*, World Scientific (2002).
[9] Drucker, H., Burges, C. J. C., Kaufman, L., Smola, A. and Vapnik, V.: Support vector regression machines, *Adv. Neural Inf. Process. Syst. 9* (Mozer, M. C., Jordan, M. I. and Petsche, T., eds.), MIT Press, Cambridge, MA, pp. 155–161 (1997).
[10] Enright, A. J., John, B., Gaul, U., Tuschl, T., Sander, C. and Marks, D. S.: MicroRNA targets in Drosophila, *Genome Biol.*, Vol. 5, No. 1, p. R1 (2004).

表 1 数値実験の結果. 2 つのパネルのそれぞれにおける第 1 行は “タンパク質の PDB ID 鎖: 相互作用部位の最初の残基の番号–最後の残基の番号” を表す. N と $\bar{l}$ はそれぞれ, 標本の大きさと正例と負例の長さの平均を表す. SVM S.K. と SVM A^* は, それぞれ 2-スペクトルカーネルを用いた SVM と本研究において構築された SVM の A^* 上のアナロジーを表す.

Table 1 Results of the simulation experiments.

	1il1 A:134–215		2bnq E:127–220		1inq B:1011–1092		1it9 H:133–216		1it9 L:123–210	
	$N = 20$	$\bar{l} = 87.20$	$N = 22$	$\bar{l} = 79.45$	$N = 32$	$\bar{l} = 83.63$	$N = 36$	$\bar{l} = 87.28$	$N = 40$	$\bar{l} = 84.05$
	SVM S.K.	SVM A^*	SVM S.K.	SVM A^*	SVM S.K.	SVM A^*	SVM S.K.	SVM A^*	SVM S.K.	SVM A^*
accuracy	0.8267	0.9433	0.9475	1.0000	0.8380	0.9140	0.8267	0.9817	0.8029	0.8900
precision	0.8833	0.9500	1.0000	1.0000	0.8860	0.9105	0.9971	0.9914	0.9527	0.8951
recall	0.7933	0.9533	0.8950	1.0000	0.8080	0.9520	0.6567	0.9733	0.6543	0.9086
F-measure	0.8359	0.9516	0.9446	1.0000	0.8452	0.9308	0.7919	0.9823	0.7758	0.9018
	1ikv A:317–419		1cff A:84–145		1iza A:1–20		1ifh L:119–206		1p2c F:1501–1627	
	$N = 46$	$\bar{l} = 171.83$	$N = 56$	$\bar{l} = 20.57$	$N = 88$	$\bar{l} = 25.02$	$N = 96$	$\bar{l} = 82.08$	$N = 126$	$\bar{l} = 111.83$
	SVM S.K.	SVM A^*	SVM S.K.	SVM A^*	SVM S.K.	SVM A^*	SVM S.K.	SVM A^*	SVM S.K.	SVM A^*
accuracy	1.0000	1.0000	0.7944	0.8211	0.8160	0.9727	0.9125	0.9738	0.9481	0.9562
precision	1.0000	1.0000	0.8698	0.8846	0.9669	0.9782	0.9663	0.9688	0.9841	0.9546
recall	1.0000	1.0000	0.7311	0.7511	0.6560	0.9693	0.8575	0.9825	0.9114	0.9610
F-measure	1.0000	1.0000	0.7944	0.8124	0.7817	0.9737	0.9087	0.9756	0.9464	0.9578

[11] Gusfield, D.: *Algorithms on Strings, Trees, and Sequences*, Cambridge University Press (1997).
[12] Haussler, D.: Convolution kernels on discrete structures, Technical Report UCSC-CRL-99-10, Department of Computer Science, University of California, Santa Cruz, Santa Cruz, CA (1999).
[13] Horton, P. B. and Kanehisa, M.: An assessment of neural network and statistical approaches for prediction of E. coli promoter sites, *Nucleic Acids Res.*, Vol. 20, No. 16, pp. 4331–4338 (1992).
[14] Jeong, H., Mason, S. P., Barabási, A.-L. and Oltvai, Z. N.: Lethality and centrality in protein networks, *Nature*, Vol. 411, No. 6833, pp. 41–42 (2001).
[15] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. and Barabási, A. L.: The large-scale organization of metabolic networks, *Nature*, Vol. 407, No. 6804, pp. 651–654 (2000).
[16] John, B., Enright, A. J., Aravin, A., Tuschl, T., Sander, C. and Marks, D. S.: Human microRNA targets, *PLoS Biol.*, Vol. 2, No. 11, pp. 1862–1879 (2004).
[17] Jordan, C.: *Cours d'analyse de l'École Polytechnique*, Gauthier-Villars, Paris (1887).
[18] Koyano, H. and Kishino, H.: Quantifying biodiversity and asymptotics for a sequence of random strings, *Phys. Rev. E*, Vol. 81, No. 6, p. 061912 (2010).
[19] Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M. and Rajewsky, N.: Combinatorial microRNA target predictions, *Nat. Genet.*, Vol. 37, No. 5, pp. 495–500 (2005).
[20] Larsen, N. I., Engelbrecht, J. and Brunak, S.: Analysis of eukaryotic promoter sequences reveals a systematically occurring CT-signal, *Nucleic Acids Res.*, Vol. 23, No. 7, pp. 1223–1230 (1995).
[21] Leslie, C., Eskin, E., Weston, J. and Noble, W. S.: Mismatch string kernels for SVM protein classification, *Adv. Neural Inf. Process. Syst. 15* (Becker, S., Thrun, S. and Obermayer, K., eds.), MIT Press, Cambridge, MA, pp. 1417–1424 (2003).
[22] Leslie, C. and Kuang, R.: Fast string kernels using inexact matching for protein sequences, *J. Mach. Learn. Res.*, Vol. 5, pp. 1435–1455 (2004).
[23] Leslie, C. S., Eskin, E. and Noble, W. S.: The spectrum kernel: A string kernel for SVM protein classification, *Proc. Pacific Symp. Biocomput.* (Altman, R. B., Dunker, A. K., Hunter, L., Klein, T. E. and Lauderdale, K., eds.), Vol. 7, pp. 566–575 (2002).
[24] Levenshtein, V. I.: Binary codes capable of correcting deletions, insertions, and reversals, *Sov. Phys. Dokl.*, Vol. 10, pp. 707–710 (1966).
[25] Lewis, B. P., Shih, I.-h., Jones-Rhoades, M. W., Bartel, D. P. and Burge, C. B.: Prediction of mammalian microRNA targets, *Cell*, Vol. 115, No. 7, pp. 787–798 (2003).
[26] Li, H. and Jiang, T.: A class of edit kernels for SVMs to predict translation initiation sites in eukaryotic mRNAs, *J. Comput. Biol.*, Vol. 12, No. 6, pp. 702–718 (2005).
[27] Lodhi, H., Shawe-Taylor, J., Cristianini, N. and Watkins, C.: Text classication using string kernel, *Adv. Neural Inf. Process. Syst. 13* (Leen, T. K., Dietterich, T. G. and Tresp, V., eds.), MIT Press, Cambridge, MA (2001).
[28] Ma, B., Elkayam, T., Wolfson, H. and Nussinov, R.: Protein-protein interactions: Structurally conserved residues distinguish between binding sites and exposed protein surfaces, *Proc. Natl. Acad. Sci. USA*, Vol. 100, No. 10, pp. 5772–5777 (2003).
[29] Mosca, R., Céol, A., Stein, A., Olivella, R. and Aloy, P.: 3did: A catalog of domain-based interactions of known three-dimensional structure, *Nucleic Acids Res.*, Vol. 42, No. D1, pp. D374–D379 (2014).
[30] Paaß, G., Leopold, E., Larson, M., Kindermann, J. and Eickeler, S.: SVM classification using sequences of phonemes and syllables, *Proc. 6th Eur. Conf. Principles Data Min. Knowl. Discov.* (Elomaa, T., Mannila, H. and Toivonen, H., eds.), Springer, pp. 373–384 (2002).
[31] Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L. and Bradley, A.: Identification of mammalian microRNA host genes and transcription units, *Genome Res.*, Vol. 14, No. 10a, pp. 1902–1910 (2004).
[32] Rose, P. W., Beran, B., Bi, C., Bluhm, W. F., Dimitropoulos, D., Goodsell, D. S., Prlić, A., Quesada, M., Quinn, G. B., Westbrook, J. D., Young, J., Yukich, B.,

Zardecki, C., Berman, H. M. and Bourne, P. E.: The RCSB Protein Data Bank: Redesigned web site and web services, *Nucleic Acids Res.*, Vol. 39, No. suppl 1, pp. D392–D401 (2011).

[33] Saigo, H., Vert, J.-P., Ueda, N. and Akutsu, T.: Protein homology detection using string alignment kernels, *Bioinformatics*, Vol. 20, No. 11, pp. 1682–1689 (2004).

[34] van Holde, K. E., Johnson, W. C. and Ho, P. S.: *Principles of Physical Biochemistry*, Prentice Hall, Upper Saddle River, NJ (2006).

[35] Vapnik, V. N.: *Statistical Learning Theory*, Wiley (1998).

[36] Vert, J.-P.: Support vector machine prediction of signal peptide cleavage site using a new class of kernels for strings, *Proc. Pacific Symp. Biocomput.* (Altman, R. B., Dunker, A. K., Hunter, L., Klein, T. E. and Lauderdale, K., eds.), Vol. 7, pp. 649–660 (2002).

[37] Vishwanathan, S. V. N. and Smola, A. J.: Fast kernels for string and tree matching, *Kernel Methods in Computational Biology* (Tsuda, K., Schölkopf, B. and Vert, J.-P., eds.), MIT Press, Cambridge, MA, pp. 113–130 (2004).

[38] Watkins, C.: Dynamic alignment kernels, Technical Report CSD-TR-98-11, Computer Science Department, University of London, Royal Holloway (1999).

[39] Yousef, M., Jung, S., Kossenkov, A. V., Showe, L. C. and Showe, M. K.: Naïve Bayes for microRNA target predictions-machine learning for microRNA targets, *Bioinformatics*, Vol. 23, No. 22, pp. 2987–2992 (2007).

[40] Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T. and Müller, K.-R.: Engineering support vector machine kernels that recognize translation initiation sites, *Bioinformatics*, Vol. 16, No. 9, pp. 799–807 (2000).