2008－BIO－12 (5)
2008／3／3

タンパク質機能情報文抽出規則の繰り返し学習における効果的学習順序の探索

竹内 正明† 宮西 一徳†† 尾崎 知伸‡ 大川 剛直†

† 神戸大学大学院工学研究科
†† 神戸大学大学院自然科学研究科
‡ 神戸大学自然科学系先端融合研究環

タンパク質の機能に関する情報は，タンパク質構造解析についての文献に記述されており，大量の文献から機能情報を自動的に抽出する技術が望まれている．本研究では，文献データの学習により生成された分類器を用いて，機能情報文を抽出し，さらに，抽出結果をユーザが評価し，次回の学習へフィードバックする繰り返し学習について検討している．本稿では，学習する文献の効果的な順番を探索することで，抽出精度の高性能化を試みた結果について述べる．

A Method to Search Effective Order of Training in Iterative Learning of Protein Function Information Extraction

Masaaki Takeuchi† Kazunori Miyanishi†† Tomonobu Ozaki‡ Takenao Ohkawa†

† Graduate School of Engineering, Kobe University
†† Graduate School of Science and Technology, Kobe University
‡ Organization of Advanced Science and Technology, Kobe University

Information on the function of the protein is described in the literature of the protein structure analysis, and the technology that automatically extracts protein function information from a lot of literature is necessary. In our research, we use a classifier in order to extract the sentences containing function information, and we are devising the iterative learning using the user's feedback.
In this paper, we propose the method that search effective order of learning literature in order to improve accuracy of classifier.

1 はじめに

タンパク質分子は他の物質と相互作用することで機能を発現することが知られており，すべての生物学的プロセスに重要な役割を果たしている[1)]．タンパク質の機能に関する情報はタンパク質構造解析実験に関する文献に記述されている．これらの文献は大量に存在するため，その中から機能情報を人手で抽出することは時間的，労力的に困難である．そこで我々は，生物学分野における専門家の支援のため，タンパク質構造解析に関する文献からタンパク質の機能に関する文を自動的に抽出するシステムに関する研究を行っている．

構造解析について記述した文献は，タンパク質構造データを蓄積するPDB[1]から参照されている．これらの文献に記述される情報には，以下のようなものが挙げられる．

1. 構造解析実験の手法についての情報
2. タンパク質の構造についての情報
3. 機能部位についての情報

本研究では機能部位に関する情報を含む文を抽出の対象とする．機能部位について記述された文（機能情報文）には，残基名，残基と相互作用する対象物質名（他のタンパク質の残基や化合物，DNAなど），相互作用名などが記述されているが，残基名や相互作用が省略されることも多く，その場合，機能情報文の抽出は容易ではない．以下に機能情報文の例を示す[2)]．

1. *"Major hydrophobic contacts occur between 230Trp61k and 409Arg93, 237Pro68k and 366Leu60, and 240Phe71k and 409Arg93."*
2. *"The two N-acetylglucosamines (NAGs) attached to 373Asn60G are shown in stereo in the electron-density map."*

[1] Protein Data Bank (http://www.rcsb.org/pdb/home/home.do)

文 1. では *"230Trp61k"*, *"409Arg93"*, *"237Pro68k"*, *"366Leu60"*, *"240Phe71k"* が残基名であり，*"hydrophobic contacts"* が相互作用名である．一方，文2.では具体的な相互作用名は出現しない．しかし，化合物 *"N-acetylglucosamines"* と残基 *"373Asn60G"* の間の *"attached to"* という部分から，これらの物質が互いに結合する可能性があり，この文が相互作用を表す機能情報文であると判断できる．

我々は，機能情報文の自動抽出を行なうために，SVM[6] を用いた繰り返し学習[4] や，次節で述べる，決定木[5] を用いた繰り返し学習を検討してきた．我々の扱う繰り返し学習では，初期学習データにより生成された分類器を用いて，未知データを分類し，分類結果をユーザが評価する．ユーザの評価によって誤分類事例を特定し，次回の学習へフィードバックを行なうものとする．つまり，未知データの一部を次回の学習に用い，そのような一連の処理を繰り返すことで，分類器の性能向上を図る．未知データの一部を学習するにあたり，学習する順番が生成される分類器に影響を及ぼす可能性がある．本稿では，生成される分類器の性能が繰り返し学習における学習順序に左右される例を示し，効果的な学習を行なうための順序を探索する手法について述べる．

2 繰り返し学習によるタンパク質機能情報文抽出方式[3]

文献を機械学習で取り扱うためにベクトル化する際，使用する属性について説明し，繰り返し学習の概要と決定木の部分的更新について述べる．

2.1 機能情報文の属性

文献を学習できる形式へ変換する際，以下の3種類の属性を用いる．なお，本研究では文献に含まれる一文を一事例として扱うため，文単位でベクトル化を行なう．

相互作用対象物質同士の原子間距離

機能情報文には相互作用する物質の組が記述されることがある．ある残基が他の物質と相互作用するとき，残基中の原子と相互作用対象物質間の距離は近接することが知られており，この性質を属性として採用する．すなわち，相互作用対象物質間の距離がある閾値よりも小さい文に対して，"1" を付与し，抽出されなかった文に対しては "0" を付与する[7]．

機能情報文に頻出する単語

機能情報を記述するときに頻繁に使用されると考えられる単語を，頻出単語として機能情報文の属性に採用する．機能情報文に出現する単語ごとの出現回数を計算し，出現回数の多いものから *"interact"*, *"bind"*, *"salt link"*, *"hydrogen bond"* など，機能情報文の属性としてふさわしいと思われる語を吟味し，属性とする．文中にこれらの単語が出現すれば "1" を，出現しなければ "0" を付与する．

パターン

本研究で対象とする文献は，既に文書中に出現する固有表現が全て特定され，固有表現タグ（<residue>，<interaction> など）が付与されていると仮定する．

属性として使用するパターンは，以下の例に示すような固有表現タグ，動詞，名詞，ワイルドカードから構成されるものである．文に対してパターンマッチングを行い，マッチした文に対して "1" を，そうでない文には "0" を付与する．

例 1 <domain>(.)*interaction(.)*<residue>

例 2 form(.)*<residue>(.)*<group>

例 3 <protein>(.)*domain(.)*<domain>

2.2 繰り返し学習の概要

機械学習の手法を用いて学習する際，学習データの事例数や属性がモデルを構築するために必ずしも十分であるとは限らない．本研究における繰り返し学習では，生成された分類器が出力した機能情報文に対して，生物学分野の専門家が「正解」「不正解」などの評価をすることを前提に，誤分類と判定された事例と，その事例から抽出したパターンを次回の学習へフィードバックする．このような，「学習」，「テスト」，「ユーザによる抽出結果の評価」を繰り返す過程で，事例や属性の数が増加させ，抽出精度向上を図る．

ここで図1に示す繰り返し学習の流れについて述べる．まず，学習データを入力とし分類器を生成する (1，2)．分類器へテストデータを入力し，抽出結果を得る (3，4)．抽出結果をシステムのユーザが評価し，誤分類事例を特定する (5，6)．誤分類結果からパターンを抽出し，そのパターンを属性としてデータに追加する (7，8，10)．また，特定された誤分類事例自体も訓練事例として学習データへ追加する (9，10)．以上の (1) から (10) の操作を繰り返す．

2.3 部分的更新を伴う繰り返し学習

繰り返し学習における分類器として決定木を用いた場合，繰り返しの過程でユーザが評価を行うことにより，決定木のノードごとの誤分類情報を

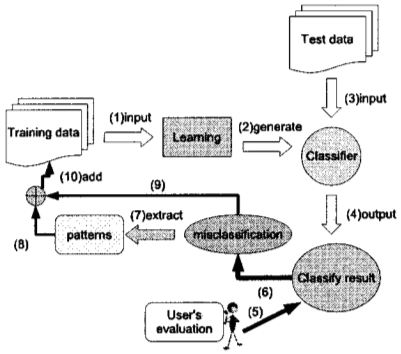

図 1: 繰り返し学習

得ることができる．誤分類率と誤分類数が閾値を超えるノードから先を，図 2 のように再構築することで，精度の良い部分木を残し，精度の悪い部分木を修正する．

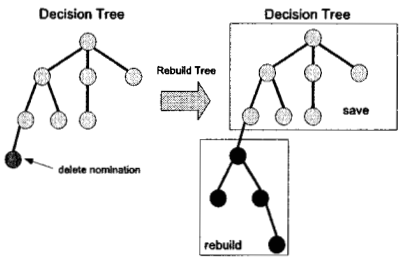

図 2: 決定木の部分的更新

3 学習順序の探索

3.1 予備実験

前節で述べた繰り返し学習において，フィードバックされる文献の順番が，分類器の性能に影響を及ぼすことを示すため，予備実験を行なった．

3.1.1 予備実験設定

実験には PDB から参照されている表 1 に示す文献を使用する．また，使用する文献には人手により固有表現タグが付与されており，機能情報文も既に特定されているものとする．

初期学習文献として “1a4j” を，評価用文献として “1a0q” を用い，表 2 のような順番でフィードバックを行なった．

3.1.2 予備実験結果

表 2 の A，B それぞれの順番における F 値の推移を図 3 に示す．

図 3 は，フィードバックの順番以外は全て同じ条件で実験を行なった結果であるが，最終的な F 値

表 1: 使用する文献

literature No.	PDB-ID	number of sentences	number of correct sentences
1	1a0h	289	26
2	1a0q	259	23
3	1a26	203	13
4	1a3l	214	23
5	1a3r	299	21
6	1a4j	190	13
7	1a5a	113	10
8	1a5h	245	39
9	1a5i	275	73
10	1a5v	241	20
11	1a5y	256	33
12	1a5z	304	8
13	2a2g	288	13
14	2a39	312	4

表 2: 予備実験におけるフィードバックの順番

A : 3 →10→8→5→7→ 9 →12→4→1→14→11→13
B : 11→14→4→9→1→10→ 5 →3→8→12→13→ 7

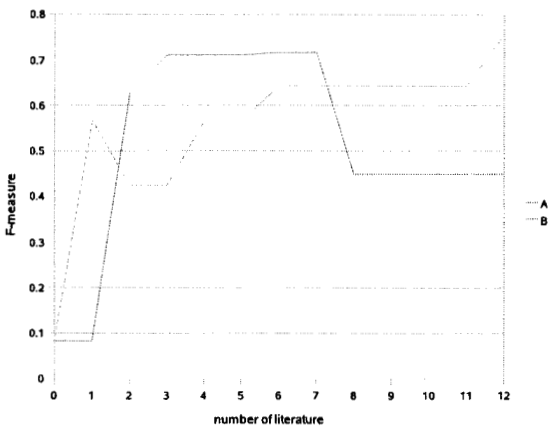

図 3: 初期学習文献 1a4j，評価文献 1a0q　F 値の推移の一例

も，その推移も大きく異なる．このことから，本研究の繰り返し学習において，学習を行なう順序が，生成される分類器の性能に影響を及ぼすことがわかる．

3.2 学習順序の探索手法

予備実験の結果から，学習順序が分類器の性能に影響を及ぼすことがわかった．ここで，繰り返し学習を行なう過程で，生成される分類器を高性

能化するための，効果的な学習順序の探索方法について述べる．

3.2.1 学習順序探索の概要

学習する順序を探索するために，図 4 のように利用可能な文献データが n 個存在する場合を考える．つまり，これまでの繰り返し学習と違い，利用可能な文献が n 個存在するときのみ，繰り返し学習を行うこととなる．

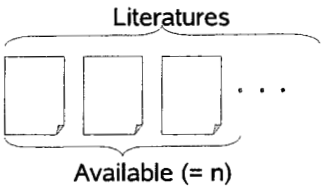

図 4: 利用可能な文献数

探索手法の流れは以下のようである．

1. n 個の文献を並び替え，在り得る全ての順序を列挙する．
2. $n!$ 通りの順番それぞれで繰り返し学習を行う．
3. $n!$ 通りの結果から，何らかの基準で最良な分類器を選択する
4. 次回，n 個の文献が利用可能となったとき，1. から同様の処理を行う．

このような流れで繰り返し学習を行うことで，n 個ごとの効果的な学習順序を発見できると考える．$n = 3$ における順序探索の例を図 5 に示す．なお，$n = 1$ のときは順序を探索しない場合と等価である．

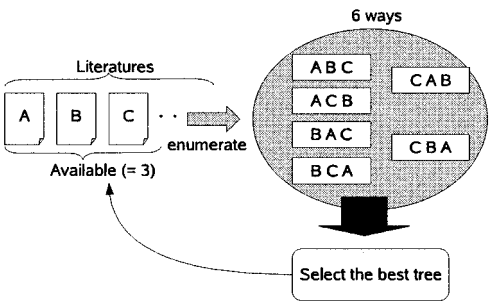

図 5: 順序探索の例

3.2.2 学習順序探索基準

最良な分類器を選択する際，各分類器の誤分類率を算出することで，それぞれの性能を比較できる．本研究では，繰り返し学習の過程で決定木のノードごとの誤分類情報を利用することができる．よって，n 個の文献を学習したときに生成された木を評価するために，$n!$ の木の中から最も誤分類率が低いものを最良な木とし，次の繰り返し学習の初期状態に採用するものとする．

4 評価及び考察

本稿で提案する学習順序の探索手法を評価する．なお，使用する文献データは予備実験に用いたものと同様で，学習順序探索時の n は 3 とする．

表 2 の A について，

(5, 1, 11) → (14, 12, 10) → (7, 6, 3) → (13, 9, 4)

のように，$n = 3$ でデータを区切り，学習順序探索を適用した結果を図 6 に示す．

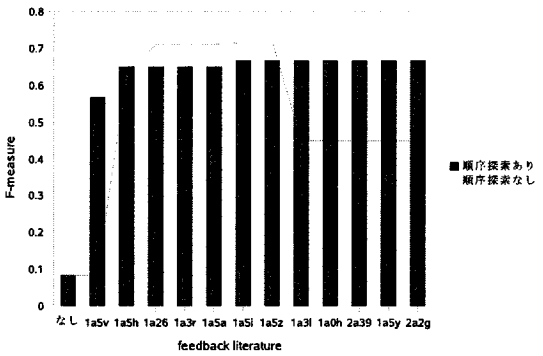

図 6: 順序 A における F 値の比較

図 6 から，順序探索を行なった結果，予備実験のときとフィードバックの順番が異なり，最終的な F 値も改善していることがわかる．図 6 の x 軸項目が順序探索によって，効果的であると判断されたフィードバック文献の順番である．

また，同様に初期学習文献に 1a4j，評価文献に 1a0q を用いた場合に，表 3 のように異なる 8 通りの順番でフィードバックを与えたときの F 値の平均を，図 7 に示す．表 3 の括弧は順序を探索する際のグループを表す．

図 7 では，最終的に到達した値は順序探索をした場合の方がやや劣るが，全体的には底上げされ，精度が改善された．

逆に，順序探索による悪影響も観測された．表 3 における，順番 C で学習を行った際の「順序探索あり」と「順序探索なし」の結果を図 8 に示す．

表 3: フィードバックの順番（初期学習文献:1a4j, 評価文献:1a0q）

A: (3 →10→ 8) → (5 → 7 → 9) → 12 → 4 → 1) →14→11→13)
B:(11→14→ 4) → (9 → 1 →10)→ 5 → 3 → 8) →12→13→ 7)
C: (1 → 4 → 5) → (7 → 8 → 9) →(10→11→12)→(3→13→14)
D: (5 → 1 →12)→ (7 → 9 →14)→ 10 →13→ 8) → 3 →11→ 4)
E:(14→13→12)→ (3 →11→ 7) → 1 →10→ 8) → 5 → 4 → 9)
F:(14→ 5 → 1) → (7 → 8 → 9) → 3 → 4 →11)→13→10→12)
G:(13→ 7 →12)→(14→10→ 4) → 3 → 9 → 8) → 1 →11→ 5)
H: (4 → 9 → 3) →(12→10→ 8) → 1 →14→ 7) → 5 →11→13)

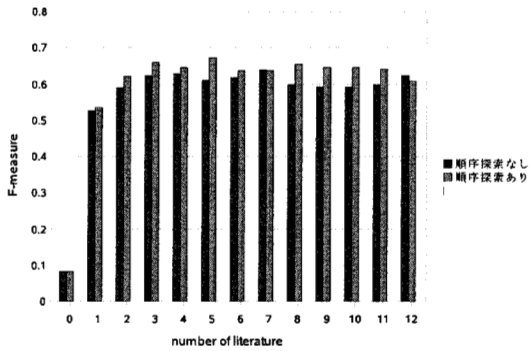

図 7: F 値の平均による比較

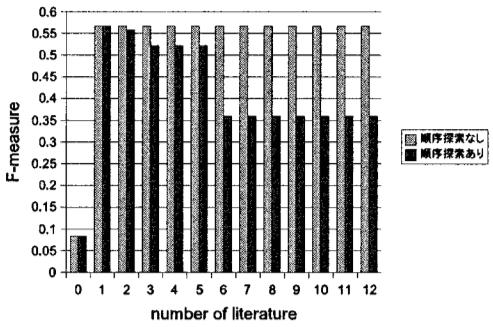

図 8: 順序探索による悪影響

順序探索を行った結果，以下のような順序が効果的として判断された．

1a3r → 1a0h → 1a3l → 1a5a → 1a5h → 1a5i → 1a5v → 1a5y → 1a5z → 1a26 → 2a2g → 2a39

最初の 3 つの文献について順序の変更が起こり，その結果悪影響が出たと思われる．

本稿で提案した手法では，利用可能な文献の数 n の中でのみ順序の入れ替えが発生する．n 個の文献に対して順序探索を行い，最初の 1 つのみを決定し，$n-1$ 個と次に利用可能となった 1 つの文献に対して，同様の順序探索を行う方法へ変更すれば，より柔軟な順序の変更が起こると考えられる．また，今回は分類器の誤分類率で順序を評価しているが，決定木の部分的更新の頻度など，別の評価基準を採用することも考えられる．

5 まとめ

ユーザの評価を利用した繰り返し学習において，学習する順序により分類器の性能が変化することを述べた．利用可能な文献が複数存在する場合を想定し，効果的に学習するための順序を，決定木の誤分類率という基準で探索する手法について述べた．評価実験の結果，F 値の平均で比較し，精度の全体的な底上げを確認することができた．

参考文献

1) J. M. Berg, J. L. Tymoczko, L. Stryer 著. 入村 達郎, 岡山 博人, 清水 孝雄 監訳, ストライヤー生化学 第 5 版. 東京化学同人, 2004.

2) P. D. Martin, M. G. Malkowski, J. Box, C. T. Esmon, and B. FP. Edwards. New insights into the regulation of the blood clotting cascade derived from the x-ray crystal structure of bovine meizothrombin des f1 in complex with ppack. *Structure*, Vol. 5, pp. 1681–1693, 1997.

3) K. Miyanishi, M. Takeuchi, T. Ozaki, and T. Ohkawa. Iterative learning with feature update for extracting sentences containing protein function information. *7th Atlantic Symposium on Computational Biology and Genome Informatics(CBGI 2007)*, 2007.

4) Md. A. Munna and T. Ohkawa. A method to extract sentences with protein functional information from literature by iterative learning of the corpus. 情報処理学会論文誌. バイオ情報学, Vol. 47, No. 17, pp. 22–30, 2006.

5) J. R. Quinlan. *C4.5: programs for machine learning*. Morgan Kaufmann Publishers Inc., 1993.

6) V. N. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York, Inc., 1995.

7) 兼田佳和, Md. A. Munna, 大川剛直. 蛋白質立体構造データを利用した文献からの蛋白質相互作用記述文抽出方式. 電気学会論文誌 C（電子・情報・システム部門誌）, Vol. 125, No. 5, pp. 690–697, 2005.