2009－BIO－16 (3)
2009／3／5

距離の再定義を伴う事例選択を用いたタンパク質機能情報文抽出方式

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文献に記述されている．しかし，大量の文献から人手で機能情報を抽出することは困難であるため，それらを自動的に抽出する技術が望まれている．機能情報の抽出を，機能情報が含まれる文の抽出と考えたとき，文献中の各文に対し，それが機能に関する文かどうかラベル付けし，機械学習を行うことで，自動抽出することが可能となるが，学習用データのラベル付けは専門家が人手で行う必要がある．ラベル付けを行なう専門家の負担を軽減するためには，学習に効果的なデータのみを検出することが有効であると考えている．本研究では，学習に効果的なデータを検出するために，ラベル付けされた文同士の特徴空間における距離の再定義を伴う事例選択手法を提案する．

A Method of Protein Function Information Sentence Extraction Using Instance Selection with Distance Metric Learning

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

Protein function information is reported in many documents. It is hard to extract the function information manually from a number of a documents. Therefore it is required to extract automatically. Extraction of protein function information is considered as to select sentences containing the information. It gives the label whether it contains the function infromation to each sentence in the literature, and it is able to extract the sentences automatically by machine learning algorithm. But experts must give labels manually to generate training data. We consider that detecting effective data for learning is useful to reduce the load of expert. In this paper, we propose instance selection method with distance metric learning to detect effective data for learning.

1 はじめに

タンパク質の機能情報は，大量に存在する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．大量の文献から，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，生物学分野の専門家支援のため，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 [1) 4)]．

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．

我々は，機能情報文の自動抽出を行なうために，決定木を用いた繰り返し学習などを検討してきた．しかし，上記の手法は初期学習文献には全てラベルが付いており，また，新規文献の抽出結果についても，全て人手でラベルが付け直されることを想定している．現実には，ラベルの付いた初期学習文献が十分存在することは，必ずしも期待できず，

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

分類結果全てについて人手でラベルを付け直していては，機能情報文抽出支援の効果が低い．そこで，ラベルの付いていない文献に含まれる一部のデータにのみ，人手でラベルを付与する状況を考慮する必要がある．本稿では，学習に効果的な事例を選択する手法と，事例選択を用いた繰り返し学習によるタンパク質機能情報文抽出システムについて述べる．

2 繰り返し学習によるタンパク質機能情報文抽出方式

本研究では，文献に含まれる 1 つの文を 1 事例として扱い，機能情報文を正例，それ以外の文を負例とする 2クラスの分類問題を考える．なお，文中に出現する固有表現は全て特定され，固有表現タグ(<residue>，<interaction> など）が付与されていると仮定する．ここでは，機械学習を行うために文を特徴ベクトルとして表現する際に，使用する属性について説明する．その後，提案手法である事例選択を組み込んだ，繰り返し学習の枠組みについて述べる．

2.1 文の属性

事例のベクトル化には，以下の 3 種類の属性を用い，48 次元のベクトルへ変換する．

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，文中に現れる相互作用対象物質間の距離がある閾値よりも小さい場合，“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．

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単語として，*“interact”*, *“bind”*, *“salt link”*, *“hydrogen bond”* などの全 45 単語を属性に採用する．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

機能情報を記述する際によく用いられる文型をパターンとして属性に採用する．属性として使用するパターンは，以下に示すようなワイルドカードを含んだものである．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．ここで，<tag> は任意の固有表現タグである．

- between(.)*<residue> and <tag>
- <residue>(.)*[動詞](.)*<tag>

2.2 事例選択を用いた繰り返し学習の概要

機械学習によって機能情報文の自動抽出を行うためには，ラベル付きデータで学習を行なう必要があるが，十分なラベル付きデータが利用可能とは限らない．提案する枠組みにおいては，少量の初期学習データを学習し，生成された分類器を用いて，ラベルの付与されていない新規文献に仮のラベルを付与する．ユーザ (専門家) が，仮ラベルを持つデータに真のラベルを付与し，再学習を行なうことで，分類精度向上を試みる．なお，分類器の学習には SVM を用いる．

事例選択を用いた繰り返し学習の流れを，図 1 に示す．まず，入力された新規文献から，ラベルを付与すべきデータを選択し (1)，選択されたデータについてユーザが真のラベルを付与する (2)．ラベル付けされた一部のデータから分類器を生成し (3)，ラベル付けされていないデータを分類器に入力し (4)，分類結果から仮のラベルを付与する (5)．そして，次回の新規文献と仮ラベル付きのデータに対して事例選択を行なう．ラベル付きデータが少量であることを考慮して，2 周目以降，仮のラベル付きデータと真のラベル付きデータ両方を用いて，学習を行うものとする (3)．以上，(1) から (6) の操作を新規文献が入力される度に繰り返す．

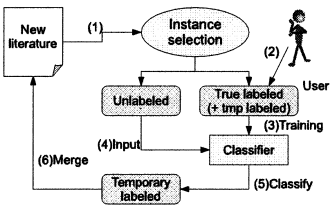

図 1: 繰り返し学習の流れ

3 距離の再定義を伴う事例選択

3.1 事例選択

事例選択については，能動学習の分野において，複数の学習アルゴリズムから生成された分類器を用いて分類を行い，分類結果の割れたデータを情報量の大きいデータとして選択する手法などが提案されている [5]．

本研究では，仮のラベルが付与された事例を学習に用いるが，仮ラベルは誤っている可能性があり，誤った仮ラベルを正すことで，学習に好影響を与えると考えられる．よって，本研究における

事例選択は，仮のラベルが誤っている可能性が高いデータを選択することと言える．

そこで，データのラベルと距離を用いて，仮ラベルの誤っている可能性を定量的に評価し，事例選択の基準とする．“距離の近いデータ同士は同じラベルを持つ可能性が高い”と考えたとき，データ x，i 間の距離を d_{xi} とすると，図2において，全てのデータのラベルが正であるとき，データ y より x の方が仮のラベルが誤っている可能性が高いと考える．逆に，データ x，y が正例で，データ 1，2 が負例であるとき，データ x より y の方が仮のラベルが誤っている可能性が高いと考える．以上の考えから，仮ラベルの誤っている可能性を次の式により定義する．

$$f(x) = \sum_i \frac{h(x,i)}{d_{xi}}, \quad h(x,i) = \begin{cases} 1(c_x = c_i) \\ -1(c_x \neq c_i) \end{cases} \tag{1}$$

ここで，c はクラスラベルとする．

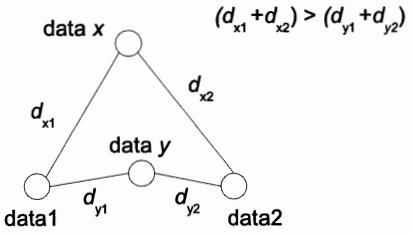

図 2: データ間の距離

仮ラベルを持つ全事例について，式(1)の値を求めたとき，値が最小の事例は，最も仮ラベルが誤っている可能性が高いと考えられる．

3.2 距離の再定義

半教師ありクラスタリングの分野では，Distance Metric Learning と呼ばれる，データ同士の制約を用いて距離を学習し，より良いクラスタを得ようとする試みがある [3]．今回の枠組みにおいて，データはラベル情報を持つため，ラベル情報を制約に変換し，距離定義に反映させていき，式(1)による事例選択の性能を向上させることを考える．Distance Metric Learning では，データ x，y 間の距離を

$$d_{xy} = ||x-y||_A = \sqrt{(x-y)^T A (x-y)}, \tag{2}$$

と定義する．ここで，A は行列である．そして，次に示すようなデータ間の制約 S，D を考える．

$$S: \ (x_i, x_j) \in S \text{ if } x_i \text{ and } x_j \text{ are similar}$$
$$D: \ (x_i, x_j) \in D \text{ if } x_i \text{ and } x_j \text{ are dissimilar}$$

これらの制約を考慮して，次の最小化問題を解くことで，式(2)における A を求め，距離の再定義を行う．

$$\min_A \sum_{(x_i,x_j)\in S} ||x_i - x_j||_A^2$$
$$s.t. \sum_{(x_i,x_j)\in D} ||x_i - x_j||_A \geq 1, \quad A \geq 0$$

本研究では，ユーザによって付与された真のラベルが付与される度に，これをもとに制約を生成し，距離の再定義を行う．これにより，真のラベルを持つ事例間で，同一ラベル同士が近づき，異なるラベル同士が離れることになり，“距離の近いデータ同士は同じラベルを持つ可能性が高い”という考えをより強く反映した特徴空間で，式(1)を用いてデータの順位付けを行うことができる．

4 評価及び考察

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．

表 1: 使用する文献

PDB ID	num of sentences	num of positives	PDB ID	num of sentences	num of positives
1a0f	382	46	1a0h	359	26
1a0k	683	19	1a0o	148	12
1a0q	295	23	1a1s	285	24
1a23	528	5	1a26	243	13
1a3a	544	17	1a3h	275	8
1a3l	272	23	1a3r	299	21
1a3s	306	7	1a3y	209	3
1a4j	190	13	1a5a	113	10
1a5h	296	39	1a5i	324	73
1a5v	277	20	1a5y	291	33
1a5z	428	8	2a2g	365	13
2a39	312	4			

表1からランダムに7つの文献を選択し，1つずつ学習用文献として入力し (図1における “New literature” にあたる)，残りの全ての文献で図1における “Classifier” を評価する．4通りの組み合わせを対象に，以下5つの手法で比較を行った結果を図3に示す．なお，4通りの結果における F 値の平均で評価を行った．

手法 1 新規文献全てにラベルを付与した場合
手法 2 提案手法 (真のラベルが付与される度に再定義した距離で式(1)を計算した場合)
手法 3 通常の距離で式(1)を計算した場合
手法 4 SVM の出力で順位付けを行う場合 [2]

手法 5 ランダムに事例を選択した場合

ここで，全ての手法において最初の事例選択はランダムに行うものとする．

本研究では，ユーザの負担を軽減することが目的であるため，20 文ずつ事例選択する場合と，さらに選択数を減らして，10 文ずつ事例選択する場合で実験を行った．図 3 が各文献から 20 文ずつ選んだ場合，図 4 が各文献から 10 文ずつ選択した場合の結果である．

提案手法は，いずれの結果においても，ランダムな事例選択である手法 5 よりも常に高い精度を保ち，また，全事例を選択する手法 1 の精度にも近い値を示している．SVM を用いた手法 4 は実験 1 において最も立ち上がりが早いが，最初に与えられる真のラベルを持つデータによっては，良好な分離平面が得られず，実験 2 のように低い精度の状態が続くことがある．一方で，式 (1) を用いた手法は比較的安定した精度を示している．提案手法 (手法 2) は，事例選択数を減らした場合でも，手法 3 と比較して，より高い精度を保っており，再定義された距離を用いた事例選択の有効性を示している．

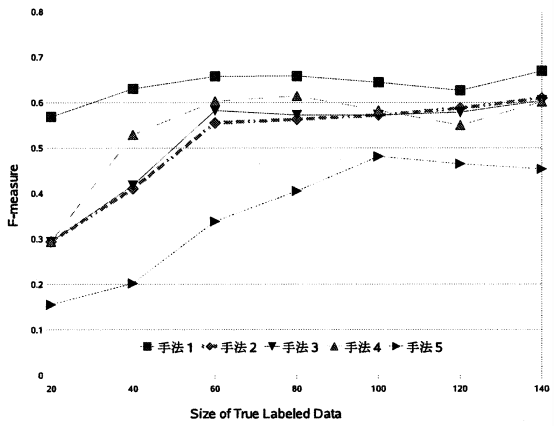

図 3: 実験結果 1

5 まとめ

本研究では，距離の再定義を伴う事例選択を用いた，タンパク質機能情報文抽出システムについて述べた．評価実験の結果，提案手法が安定した精度を保ち，少数の事例選択においても，高い精度を示すことを確認した．今後の課題として，本研究では，再定義された距離を事例選択時にのみ利用したが，SVM のカーネル関数に学習された距離を組み込むことで，分類時にも利用することが挙げられる．

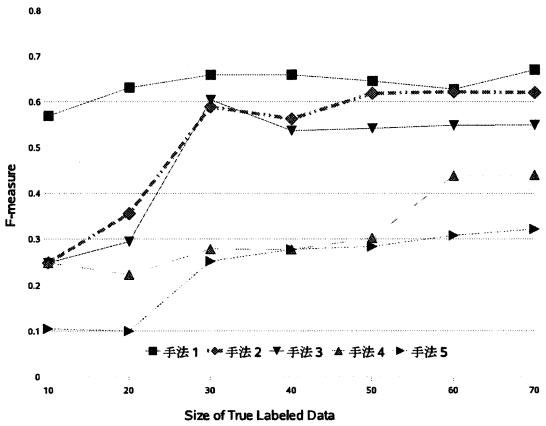

図 4: 実験結果 2

参考文献

1) K. Miyanishi, M. Takeuchi, T. Ozaki, and T. Ohkawa. Iterative learning with feature update for extracting sentences containing protein function information. In *Proceedings of 7th Atlantic Symposium on Computational Biology and Genome Informatics(CBGI 2007)*, pp. 96–102, 2007.

2) M. Sassano. An empirical study of active learning with support vector machines for japanese word segmentation. In *Proceedings of 40th Annual Meeting on Association for Computational Linguistics(ACL 2002)*, pp. 505–512, 2001.

3) E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Proceedings of Advances in Neural Information Processing Systems 15*, Vol. 15, pp. 505–512, 2003.

4) 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 , Vol. 125, No. 5, pp. 690–697, 2005.

5) 拓馬見塚, 直樹安倍. 集団能動学習 : データマイニング・バイオインフォマティックスへの展開 (情報論的学習理論論文小特集). 電子情報通信学会論文誌, Vol. 85, No. 5, pp. 717–724, 2002.