LE-001

生物医学文献での蛋白質名認識における過学習と Transductive SVM を用いた過学習の軽減
Overfitting in protein name recognition on biomedical literature and method of preventing it through use of transductive SVM

村田 真樹† 三森 智裕‡§ 土井 晃一*
Murata Masaki Tomohiro Mitsumori Kouichi Doi

1. はじめに

近年，機械学習の方法が蛋白質名認識の研究に使われるようになった．しかし，ある分野で学習した分類器は過学習し，その分野以外で用いると性能が大きく下がるという問題がある．そこで，我々は，breast cancer について記述している新しいコーパスを作り，GENIA コーパスで学習した分類器を用いてそのコーパスで蛋白質名認識の実験を行い，異なる分野のコーパスを学習に用いた場合の性能低下を調べた．過学習の問題を避けるために，我々は Transductive SVM を利用した．Transductive SVM は学習の際に学習データだけでなくテストデータも利用するため，過学習の問題を解決することに役立つと考えて利用した．実験において，Transductive SVM が過学習の問題を解決し，通常の SVM よりも高い精度を出すことを確認した（sub の評価基準で二種類の実験でそれぞれ F 値が 75.46%が 79.64%に 70.63%が 74.61%に向上した）．

2. 先行研究と本研究の位置づけ

近年，生物医学文献から情報を取り出すために自然言語処理技術を利用した研究が数多く報告されている．典型的な研究としては，蛋白質の相互作用や蛋白質の機能を取り出す研究があげられる．この抽出の最初の段階は専門用語としての蛋白質の抽出である．

蛋白質自動抽出では種々の手法が提案されている．Fukuda ら[1]と Franzen ら[2]は人手で作成したルールに基づいて要約文中の蛋白質名を自動抽出した．ルールに基づく方法では新しい蛋白質名が出現した時にそれに応じて新しいルールを作成する必要がありルールの管理が複雑になる．

他の方法としては辞書に基づく方法がある．この方法では，蛋白質名とそれに対応する遺伝子名が格納されている SWISS-PROT[3]のようなデータベースを利用して，パターンマッチングや動的計画法に基づいて蛋白質名の認識を行う．しかし，この方法では，同じ表記で蛋白質名であるものと蛋白質名でないものがある場合に，蛋白質名でないものを蛋白質名と誤ってしまう問題がある．例えば，"Maltose-binding protein"は MAP と略されるが，MAP は地図を意味する単語でもある．また，辞書に基づく方法の他の欠点としては辞書に登録されていない蛋白質名を抽出できないという問題がある．

生物医学文献の用語抽出に機械学習を利用する研究がなされてきた[4,5,6]．機械学習を用いる方法では，学習データに正解の情報を付与する必要がある．Collierら[4] は 100 個の要約に正解情報のタグ付けを行い，隠れマルコフモデルにより専門用語認識を行った．風間ら[5]はGENIAコーパスを学習データとして用いてSVMを用いた専門用語認識を行った．Biocreative[1] ワークショップのタスク 1 では遺伝子名，蛋白質名自動認識のコンテストがあった．そこでは機械学習を用いる方法が有効であった．しかし，この機械学習を用いる方法は，機械学習に必要な学習データを作成するコストが大きい[7]．GENIAコーパスは"human"，"blood cell"，"transcription factor"の三語を含む要約を集めたものであり，このコーパスを用いた機械学習に基づく蛋白質名自動認識の精度はおよそ 80%である．

生物医学文献から情報抽出を行う研究者にとってはある特定の分野の情報ではなく広い分野の情報を取り出すことが理想的である．広い分野の学習データを作成することは多大なコストを要する．本研究の目的は，ある特定の分野で学習した分類器の柔軟性を調査することである．ある特定の分野の蛋白質名を認識するように学習した分類器がどのくらいの性能で異なる分野で蛋白質名を認識できるかを調べた．ある特定の分野で学習した分類器は過学習を起こすことが知られている[8,9]．過学習とは，特定の分野で学習した分類器はその分野に特化しすぎて他の分野に用いる場合に性能が大幅に低下する問題である．我々は実験により生物医学分野の蛋白質名認識の問題において過学習が起きることを確認した．さらに Transductive SVM を利用することで過学習を軽減できることを確認した．

3. 方法（SVM と Transductive SVM）

サポートベクターマシン(SVM)[10]は与えられた学習データでの二つの分類の間の間隔（マージン）を最大にする超平面を求めてそれを分類の境界として利用する方法である．線形の分類が難しいときはカーネルトリックを用いて高次の非線形の分類をする．本研究では以下の多項式カーネルを用いる．

$$K(x_i, x_j) = (x_i \cdot x_j + 1)^d \qquad (1)$$

d はカーネルの次数を指す．

Transductive SVM[11]の概要は以下のとおりである．

† 独立行政法人 情報通信研究機構
murata@nict.go.jp
National Institute of Information and Communications Technology.

‡ 奈良先端科学技術大学院大学
mitsumori01@yahoo.co.jp
Nara Institute of Science and Technology.

§ 現在，宮園特許事務所

* ファーマセキュリティコンサルティング
doy@pharmasecurity.jp
Pharma Security Consulting Inc..

[1] http://www.pdg.cnb.uam.es/BioLINK/BioCreative.eval.html

1. 境界の超平面をラベルつきデータ（学習データ）から求める．
2. その超平面を利用してラベルなしデータ（テストデータ）を分類する．
3. ステップ 2 で正例または負例と判断されたデータから超平面の近くのデータペアを選ぶ．
4. ステップ 3 で選ばれたペアのラベルを交換する．ただしこの交換によりマージンが小さくなる場合は交換しない．
5. 終了条件を満足すると終了し，そうでない場合はステップ 2 に戻る．

4. 実験

4. 1 タグ付きコーパス

この研究では GENIA コーパスと breast cancer (乳がん)に関するコーパスを用いた．GENIA コーパスは生物医学テキストから情報を取り出すために自然言語処理技術を利用しやすくすることを目的に作成された．MEDLINE 中の “human”，“blood cell”，“transcription factor”という三つの文字列が共起する 2,000 要約からなる．生物医学分野の意味的なタグが付与されている．蛋白質名として protein_complex，protein_family_or_group など，いくつかのサブクラスが定義されている．本研究ではこれらのサブクラスをすべて蛋白質名として扱った．我々は，breast cancer に関係するコーパスを作成した．breast cancer に関する要約であり，かつ，蛋白質名を含む 1,000 要約を MEDLINE から抜き出し作成した．このコーパスでの蛋白質名は，我々の指導の下で生物学の専門家がタグ付けを行った．蛋白質名の定義は GENIA コーパスに従った．

4. 2 素性，パラメータ，評価方法

本研究の実験で用いられた素性を表 1 に示す．これらの素性は生物医学文献での蛋白質名抽出に従来から用いられているものである[4][12]．単語と品詞の素性は従来の自然言語処理においても用いられているものである．記号素性は，蛋白質名が大文字やギリシャ文字や数字を含むものがあるため，蛋白質名抽出に効果的であり，蛋白質抽出によく使われるものである．末尾文字列素性は，蛋白質名が ~ase や~in などの文字で終わることが多いために蛋白質名抽出に有効であり，よく使われるものである．先頭文字列素性は，効果はやや小さいがそれでも有効であることが確認されている[12]．素性抽出の例を図 1 に示す．図では単語を「単語」の列に縦方向に記述している．抽出された素性は他の列に記述している． T，O は後述する TO タグである．現在処理中の単語(“B”)の素性には，灰色で網掛けした部分のものが用いられる．前方二単語，後方二単語の情報も素性に用いられる．すでに解析済みの前方二単語の分類(T,O)も素性に利用する．SVM には線形カーネルと多項式カーネルを利用した．蛋白質認定は，文頭から文末の方向に処理した．本研究では TO タグを利用した．TO タグは T タグが蛋白質名の単語に付与され，O タグが蛋白質名でない単語に付与される．

本実験で用いた評価基準を表 2 に示す．蛋白質名は複合語の場合がある．蛋白質名が 1 個の単語である場合は，正解と不正解の評価結果だけである．蛋白質名が 2 個以上の単語からなる場合は部分的な一致も考慮にいれた評価基準を設けた．exact は最も厳格な評価基準であり，left と right は少し緩めの基準である．sub は最も緩い基準である．

表1 SVM の学習に用いる素性

素性	説明
単語	学習データに出現した単語
品詞	Brill tagger によりもとめた単語の品詞
記号	12, alpha, -(ハイフン)等の記号であるか
先頭文字列	1 から 3 の単語の先頭文字列
末尾文字列	1 から 3 の単語の末尾文字列

	単語	品詞	記号	先頭文字列	末尾文字列	分類
位置 -3	such	JJ	小文字	s su suc	h ch uch	O
位置 -2	as	IN	小文字	a as -	s as -	O
位置 -1	NF-kappa	NNP	ギリシャ文字	N NF NF-	a pa ppa	T
位置 0	B	NNP	大文字一文字	B - -	B - -	(T)
位置 +1	that	IN	小文字	t th tha	t at hat	O
位置 +2	are	VBP	小文字	a ar are	e re are	O
位置 +3	constitutively	RB	小文字	c co con	y ly ely	O

図 1 “… such as NF-kappa B that are constitutively …”から素性抽出した例

表 2 本実験で用いた評価基準

評価基準	説明
extract	システムで求めた蛋白質表現が正解と完全に一致した場合に正解とする．
left	システムで求めた蛋白質表現が正解と末尾の終了位置が異なっていても表現の先頭の開始位置が一致すれば正解とする．
right	システムで求めた蛋白質表現が正解と先頭の開始位置が異なっていても表現の末尾の終了位置が一致すれば正解とする．
sub	システムで求めた蛋白質表現が正解と一部でも一致すれば正解とする．

4. 3 同一分野での実験

学習データとテストデータが同一の分野のものを用いて実験を行った．蛋白質名認定の評価には適合率と再現率と F 値を用いた．適合率はシステムの出力した蛋白質表現のうちどれだけ正解していたかを示す．再現率は本来の蛋白質表現のうちシステムがどれだけ正しく出力できたかを示す．F 値は適合率と再現率の調和平均である．GENIA コーパスでの実験結果を表 3 に示す．10 分割の交叉検定で実験した．この実験では 2 次の多項式カーネルを利用した．2 次の多項式カーネルが最もよいことは先行研究[12]で確かめている．breast cancer コーパスでの蛋白質名認定の実験結果を表 4 に示す．これらの実験では，適合率が再現率よりも高かった．breast cancer コーパスでの結果は，GENIA コーパスの結果よりも良かった．これは，breast cancer コーパスに出現した蛋白質名の種類の数が GENIA

表 3　GENIA コーパスで通常の SVM を用いた場合の実験結果

評価基準	適合率(%)	再現率(%)	F 値(%)
extract	79.08	72.57	75.68
left	83.96	77.05	80.34
right	85.89	78.82	82.19
sub	89.42	83.41	86.30

表 4　breast cancer コーパスで通常の SVM を用いた場合の実験結果

評価基準	適合率(%)	再現率(%)	F 値(%)
extract	85.54	80.18	82.74
left	88.85	83.29	85.95
right	88.08	82.57	85.21
sub	90.90	85.53	88.10

表 5　GENIA コーパスを学習データ，breast cancer コーパスをテストデータに用いた場合の通常の SVM の実験結果

カーネル	評価基準	適合率(%)	再現率(%)	F 値(%)
線形	exact	66.30	53.26	59.07
	left	78.63	63.16	70.05
	right	71.38	57.33	63.59
	sub	82.98	68.86	**75.26**
多項式 (2 次)	exact	69.08	53.80	**60.49**
	left	80.50	62.69	**70.49**
	right	73.54	57.27	**64.39**
	sub	84.78	67.12	74.92
多項式 (3 次)	exact	68.87	47.10	55.94
	left	80.63	55.14	65.49
	right	73.09	49.98	59.36
	sub	84.87	58.70	69.40

コーパスに出現した蛋白質名の種類の数よりも小さかったためと思われる．

4．4　異なる分野のコーパスを利用した実験

ある特定の分野で学習した学習器を他の分野の蛋白質名認定が利用できるかを確かめるために，学習データとテストデータで異なる分野のコーパスを利用した場合の実験を行った．

まず，通常の SVM を用いた実験を行った．

GENIA コーパスを学習データに用い，breast cancer コーパスをテストデータに用いて通常の SVM を使って蛋白質名認定の実験を行った．その結果を表 5 に示す．breast cancer コーパスを学習データに用い，GENIA コーパスをテストデータに用いて通常の SVM を使って蛋白質名認定の実験を行った結果を表 6 に示す．これらの実験では線形カーネルと 2 次，3 次の多項式カーネルを利用した．これらの結果は表 3，表 4 の結果に比べて明らかに悪い．

次に Transductive SVM を利用した実験を行った．

GENIA コーパスを学習データに用い，breast cancer コーパスをテストデータに用いて Transductive SVM を使って蛋白質名認定の実験を行った．その結果を表 7 に示す．

表 6　breast cancer コーパスを学習データ，GENIA コーパスをテストデータに用いた場合の通常の SVM の実験結果

カーネル	評価基準	適合率(%)	再現率(%)	F 値(%)
線形	exact	62.11	53.47	57.47
	left	70.46	60.65	65.19
	right	66.45	57.20	61.48
	sub	76.43	65.65	**70.63**
多項式 (2 次)	exact	64.03	53.12	**58.06**
	left	72.05	59.77	**65.34**
	right	67.93	56.35	**61.60**
	sub	77.67	64.13	70.25
多項式 (3 次)	exact	63.66	46.88	54.00
	left	72.39	53.31	61.40
	right	67.14	49.44	56.95
	sub	77.76	56.99	65.77

表 7　GENIA コーパスを学習データ，breast cancer コーパスをテストデータに用いた場合の Transductive SVM の実験結果

カーネル	評価基準	適合率(%)	再現率(%)	F 値(%)
線形	exact	60.65	60.20	60.42
	left	72.50	71.95	72.22
	right	65.34	64.86	65.10
	sub	76.86	78.34	77.59
多項式 (2 次)	exact	56.32	75.26	**64.43**
	left	64.09	85.64	**73.31**
	right	61.80	82.57	**70.69**
	sub	69.11	93.96	**79.64**
多項式 (3 次)	exact	55.54	75.04	63.83
	left	63.11	85.28	72.53
	right	60.67	81.98	69.73
	sub	68.50	93.24	78.98

breast cancer コーパスを学習データに用い，GENIA コーパスをテストデータに用いて Transductive SVM を使って蛋白質名認定の実験を行った結果を表 8 に示す．これらの実験でも線形カーネルと 2 次，3 次の多項式カーネルを利用した．この結果は通常の SVM を利用した場合(表 5,6)よりも良かった．通常の SVM は再現率よりも適合率が良いが，Transductive SVM は適合率と再現率が同じくらいか再現率の方が大きかった．F 値は Transductive SVM は通常の SVM よりも良かった．

5．おわりに

ある特定の分野で学習した蛋白質名認識の学習器が他の分野でどのくらい利用できるかの実験を行った．本研究では，GENIA コーパスと breast cancer コーパスの二種類のコーパスを用いた．学習データとテストデータとして同一分野のものを用いた場合，GENIA コーパスで 75.68%，breast cancer コーパスで 82.74%の F 値を得た．次に，学習データとテストデータとして異なる分野のものを用いた実験を行った．GENIA コーパスで学習し，breast cancer コーパスで蛋白質名抽出を行った場合 60.49%の F 値を，breast cancer コーパスで学習し，GENIA コーパスで蛋白質抽出を

表 8　breast cancer コーパスを学習データ，GENIA コーパスをテストデータに用いた場合の Transductive SVM の実験結果

カーネル	評価基準	適合率(%)	再現率(%)	F 値(%)
線形	exact	54.74	58.86	56.73
	left	63.98	68.80	66.31
	right	60.25	64.78	**62.43**
	sub	72.63	76.71	**74.61**
多項式 (2 次)	exact	56.97	57.47	**57.22**
	left	66.34	66.92	**66.63**
	right	62.11	62.66	62.38
	sub	74.62	73.89	74.25
多項式 (3 次)	exact	54.80	55.42	55.11
	left	64.94	65.68	65.31
	right	59.59	60.26	59.93
	sub	72.85	72.27	72.56

抽出を行った場合 58.06%の F 値を得た．大きな精度低下であった．この結果は異なる分野のデータで学習した学習器は他の分野ではあまり役立たないことを意味する．異なる分野で学習した学習器の性能を改善するために我々は Transductive SVM を利用した．Transductive SVM は学習の際に学習データだけでなくテストデータも利用するため，過学習の問題を解決することに役立つと考えて利用した．Transductive SVM の利用により，GENIA コーパスで学習し，breast cancer コーパスで蛋白質名抽出を行った場合，F 値は 60.49%から 64.43%に改善した．breast cancer コーパスで学習し，GENIA コーパスで蛋白質名抽出を行った場合，F 値は，通常の SVM が 58.06%，Transductive SVM が 57.22%であった．しかし部分一致などの評価基準では Transductive SVM の有効性が確認できた．Transductive SVM の利用により left の評価基準では F 値は 65.34%から 66.63%に，right の評価基準では 61.60%から 62.38%に，sub の評価基準では 70.25%から 74.25%に改善できることを確認した．蛋白質名の境界はタグ付与者によって様々に変わることが知られている[13]．例えば，“human NF-kappa B protein”の単語列に対し“NF-kappa B”を蛋白質名とタグ付けしたり，“human NF-kappa B”や“NF-kappa B protein”や“human NK-kappa B protein”を蛋白質名とタグ付けしたりするなど様々である．ほとんど同じ意味のものであってもこのように様々にタグ付けされる．蛋白質名の境界があいまいであることを考慮すれば，left，right，sub の評価で確認できた Transductive SVM の効果は意味があると考える．実験では通常の SVM では適合率が再現率よりも高かった．Transductive SVM を用いた場合は適合率と再現率はほとんど同じであった．特定の分野で学習した分類器が過学習の問題を引き起こすことを示し，またその過学習を Transductive SVM を利用することで過学習を軽減できることを確認した．しかし，Transductive SVM は通常の SVM に比べて低い適合率と，高い再現率，F 値を得た．もしユーザにとって適合率が重要である場合は通常の SVM を利用するべきである．しかし，新しい蛋白質が発見され続け，新しい文献から新しい蛋白質名を高い適合率と F 値で抽出したい場合は，Transductive SVM を利用するのがよい．

参考文献

[1] K. Fukuda, A. Tamura, T. Tsunoda, and T. Takagi, Toward information extraction: Identifying protein names from biological papers, Proceedings of the Pacific Symposium on Biocomputing, (1999), pp. 705-716.

[2] Kristofer Franz'en, Gunnaar Eriksson, Fredric Olssonand Lars Asker, Per Lid'en, and Joakim C"oster, Protein names and how to find them, International Journal of Medical Informatics, Vol. 67, (2002), pp. 49-61.

[3] Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J. Martin, Karine Michoud, Claire O'Donovan, Isabella Phan, Sandrine Pilbout, and Michel Schneider, The swiss-prot protein knowledgebase and its supplement TeEMBL in 2003, Nucleic Acids Research, Vol. 31, No. 1, (2003), pp. 365-370.

[4] Nigel Collier, Chikashi Nobata, and Jun-ichi Tsujii, Extracting the names of genes and gene production with a hidden Markov model, Proceedings of the 18th International Conference on Computational Linguistics (COLING'2000), (2000), pp. 201-207.

[5] Jun'ichi Kazama, Takaki Makino, Yoshihiro Ohta, and Jun'ichi Tsujii, Tuning support vector machines for biomedical named entity recognition, In the proceedings of the Natural Language Processing in the Biomedical Domain (ACL2002), (2002), pp. 1-8.

[6] Ping Chen and Hisham Al-Mubaid, Context-based term disambiguation in biomedical literature, In the proceedings of the 19th International FLAIRS Conference, (2006).

[7] Jason Baldridge and Miles Osborne, Active Learning and the Total Cost of Annotation, Proceedings of Empirical Method for Natural Language Processing, (2004), pp. 9-16.

[8] Takeshi Masuyama and Hiroshi Nakagawa, Two step pos selection for svm based text categorization, IEICE Transactions on Information and Systems, Vol. E87-D, No. 2, (2004), pp. 373-379.

[9] Joseph Drish, Obtaining calibrated probability estimates from Support Vector Machines, Final Project for CSE 254: Seminar on Learning Algorithms, (2001).

[10] Vladimir N. Vapnik, The Nature of Statistical Learning Theory, (Springer., 1995).

[11] Thorsten Joachims, Transductive inference for text classification using support vector machines, International Conference on Machine Learning (ICML), (1999), pp.200-209.

[12] Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi, and Hirohumi Doi, Gene/protein name recognition based on support vector machine using dictionary as features, BMC Bioinformatics, Vol. 5, No. Suppl 1, (2005), p.S8.

[13] Tomohiro Mitsumori, Sevrani Fation, Masaki Murata, Kouichi Doi, and Hirohumi Doi, Boundary correction of protein names adapting heuristic rules, Proceedings of Computational Linguistics and Intelligent Text Processing (CICLing 2004), (2004), pp. 172-175.