3ZH-4

論文のアブストラクトを用いることによる生命医学論文の分類

吉田　貴哉†　　大和田　勇人†
†東京理科大学　理工学部　経営工学科

1　はじめに

近年，生命情報工学の分野において，発表される論文数は年々増加している．それに伴い，論文データベースから知識を抽出し登録しているデータベースも数多く存在する．しかし，人手による登録ではカバーしきれなくなってきている現状がある．そのような現状から，論文データベースから自動で知識を抽出する技術への期待が高まっている．そのようなシステムを考えるにあたって，与えられた論文が知りたい情報と関連しているのか否かを判断する技術が必要である．

また，生命情報工学の分野でプロテオームと呼ばれるタンパク質研究が注目されている．タンパク質の詳しい構造解析や機能解明は，病気の治療や創薬にも直結すると考えられるためである．

そこで本研究では，生命医学論文データベース PubMed の論文が，タンパク質相互作用に関連があるか否かをそのアブストラクトから判定する手法を提案する．

2　関連研究

W.J.Wilbur らは，PubMed に掲載されている論文のアブストラクトに対し，既知のデータベースからタンパク質名等の同定を行い，バイグラムなどの様々な単語や文法の特徴を素性とした．さらに，フーバーのロバスト学習とサポートベクターマシン (SVM) を用いることで，その論文がタンパク質相互作用に関する内容か否かを判定する手法を提案した [1]．しかし，既知のデータベースから同定を行う方法では，未知のタンパク質やまだ登録が済んでいない新しいタンパク質には対応できないという問題点が考えられる．

また，K.Fukuda らはルールベースで文書中のタンパク質名の同定を行う手法を提案している [2].

そこで本研究では，論文のアブストラクトに対して，既存のデータベースを用いるのではなくルールベースでタンパク質名等の同定を行い，特徴ベクトルを作成し SVM を用いて学習し分類することで，論文がタンパク質相互作用に関係するか否か判定する手法を提案する．

Using Abstracts to Classify Life Medicine Papers
†Takaya YOSHIDA　†Hayato OHWADA
†Department of Industrial Administration,Faculty Of Science and Technology,Tokyo University of Science

3　提案手法

SVM を用いるには特徴ベクトルの作成が必要である．以下に本研究での特徴ベクトル作成の手順を述べる．

3.1　BOW(Bag Of Words)

まず，論文のアブストラクトの全単語の中で記号や数字のみで構成されるもの，”he ”などの頻出単語及び ”(”のような記号であるものをとりのぞく．そして残った単語に対しルールベースのタンパク質名の同定を行う．この手法に関しては後述する．同定を行った後，すべての単語を小文字化し，品詞のタグ付けを行う．さらに，品詞が動詞，または名詞であるもののみを抽出する．そして抽出したものに対し，ポーターのステミングを行い，それらのうち，全文書中で 3 回以上出現するものを Bag Of Words(BOW) とする．BOW の各単語に対し，もし対象文書中にその単語が存在するなら 1 を，存在しないのなら 0 を与える．

また，Friedman によって作成されたタンパク質相互作用に関係があると考えられるキーワードも用いる [3]. このキーワードを小文字化，ステミングし，もしキーワードが文書中に 7 回以上存在するのなら新たな属性値として 1 を，そうでなければ 0 を与える．

こうして，BOW の単語の種類数より 1 だけ大きい次元数を持つ特徴ベクトルを各論文に対し作成する．

3.2　タンパク質名の同定

タンパク質相互作用に関係があるか否かを判定するのが目的であって，そのタンパク質が何かということは関係がないので，タンパク質名を同定し，全て同じタンパク質であるとみなすことは有効であると考えられる．そうすることで次元数の削減も行うことが出来る．

また，タンパク質名は，D1-Cdk4 ”のように通常の単語とは異なった特徴を持っている．よって，タンパク質名をルールベースで同定することができると考えられる．本研究では以下の手順で同定を行う．

- STEP1
  大文字のみまたは小文字のみで構成される単語は取り除く．

- STEP2
  残った単語のうち，数字が含まれるもの，またはハイフンと小文字で構成され，かつ長さが8以下の単語はタンパク質名とみなし，”p-key ”という単語に置き換える．

文書分類問題では BOW を用いると特徴ベクトルの次元数が膨大になってしまうという問題がある．しかしこの同定の処理を行うことで次元の圧縮も行うことができる．こうして作成した特徴ベクトルに対し，SVM を用いることで学習，分類を行う．

4 実験

実験には，BioCreAtIvE　の Article Classification Task (ACT) のデータセットを用いた．BioCreAtIvE とは，生命医学文献への情報抽出，テキストマイニングを発展させることを目標とした取り組みである．また，ACT とは PubMed の論文がタンパク質相互作用に関係しているか否かを自動で判定することを目的としたタスクである．ACT の各データは PubMed のアブストラクトであり，これらのデータは専門家の手によって分類済みである．

今回はトレーニングセットと開発セット，および BioCreAtIvE　の同様の目的のタスクのトレーニングセット，計 11775 のデータを用い学習を行い，各論文に対して特徴ベクトルを作成し，SVM で学習してモデルを作成した．そして，テストセット 6000 データに対しても同様の特徴ベクトルを作成し，SVM で分類し，その効果を検証した．

また，SVM のソフトウェアは数多く存在するが，本研究では SVM light を用いた．

5 評価と考察

表 1 に分類の結果を示す．ここで，1 は正 (タンパク質相互作用に関係がある)，0 は負 (タンパク質相互作用に関係がない) である．実際の BioCreAtIvE　で，一番良い結果を示したチームの手法と，本手法を比較したものを表 2 に示す．本手法は既存の手法と比較して

表 1: 分類結果

		実際		
		1	0	計
予測	1	556	490	1046
	0	354	4600	4954
	計	910	5090	

表 2: 結果の比較

	提案手法	既存手法
specificity	0.904	0.943
accuracy	0.859	0.889
sensitivity	0.611	0.585
F-score	0.569	0.614

再現率以外は上回っていない．その理由の一つとして，ルールベースでのタンパク質名の同定による精度の低下が考えられる．しかし，ルールベースで行ったことによって，タンパク質データベースなどへの登録がまだ済んでいない新しいタンパク質に関する論文に対しても同様の性能を発揮することが期待できる．

また，再現率が高いので既存手法よりも，関係のある論文を見逃さないということが言える．結果として，完全に自動分類出来るほどの性能ではないが，現状では研究者を助ける補助的なシステムとしての役割が期待できる．

6 まとめ

本研究では，論文がタンパク質相互作用に関係あるのか否かをアブストラクトから判定することを目的とした．そしてルールベースのタンパク質名同定と BOW，SVM を用い，BioCreAtIvE のデータセットに対し学習，分類を行った．その結果，既存の手法と比べて再現率は上回る結果となった．

また，本研究はタンパク質名の同定をルールベースで行うので，データベースからの同定を行う既存の手法と比べてより汎用的な手法であるということが言える．

参考文献

[1] S.Kim,W.J.Wilbur：”Classifying protein-protein interaction articles using word and syntactic features” BMC Bioinformatics 2011,12(Suppl 8),S9

[2] K.Fukuda,T.Tsunoda,A.Tamura,T.Takagi：”Toward Information Extraction:Identifying protein names from biological papers”Pac Symp Biocomput,19998,707-18

[3] C.Friedman,P.Kra,H.Yu,M.Krauthammer,A.Rzhetsky：”GENIES:a natural-language processing system for the extraction of molecular pathways from journal articles” Bioinformatics,17(Suppl. 1),S74-S82