G-001

アブストラクトを用いたタンパク質相互作用に関する医学生物学文献の分類
Classification of biomedical literature on the interaction between proteins using the abstract

吉田　貴哉 †　　金盛　克俊 ‡　　大和田　勇人 ‡
Takaya Yoshida　　Katsutoshi Kanamori　　Hayato Ohwada

1　序論

近年，医学生物学の分野において，PubMed などの文献データベースに発表される文献数は年々増加している．そのような現状から，文献から自動で様々な知識を抽出する技術への期待が高まっている．知識抽出を行うにあたってまず必要なのが，抽出したい知識にその文献が関連しているか否かの分類を行う技術である．

2　既存研究およびその問題点

Wilbur らは，PubMed に掲載されている文献のアブストラクトに対し，既知のデータベースからキーワードの同定を行った上で，文献がタンパク質相互作用に関連があるか否かを分類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．この問題の解決法として，ルールベースを用いた同定がある [2]．そこで本研究では，文献のアブストラクトに対して，既存のデータベースではなくルールベースを用いてタンパク質名の同定を行い，特徴ベクトルを作成し SVM を用いて学習し分類を行う手法を提案する．同定を行うことで特徴ベクトルの次元数の削減ができると考えられる．

3　提案手法

提案手法全体の流れを図 1 に示す．図からわかるように，本研究ではトレーニングデータとタンパク質相互作用キーワードから特徴ベクトルを作成し，テストデータから同様に作成した特徴ベクトルと SVM(Support Vector Machine) を用いることで分類を行う．

3.1　BOW(Bag Of Words)

まず，文献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”などの頻出単語をとりのぞく．これらの単語は重要ではないと考えられるからである．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．同定の処理により分類を行いやすくする．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い，副詞などを削除する．そして残ったものに対し，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．こうして作成した BOW の各単語に対し，本研究ではもし対象文献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 を与えて特徴ベクトルを各文献に対し作成する．

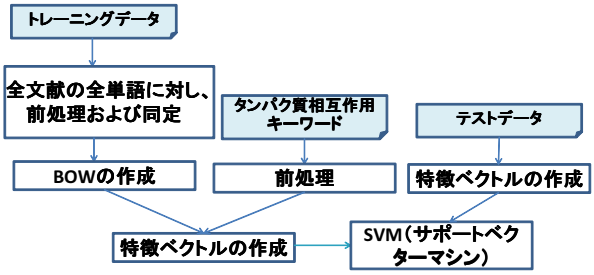

図 1: 提案手法全体の流れ

3.2　同定の方法

今回，タンパク質相互作用に関連があるか否かの分類が目的なので個々のタンパク質名は重要ではないと考えられる．また，タンパク質名は，“c-jun”や “ 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有効であると考えられる．さらに，BOW を文書分類問題に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．しかしこのルールベースでの同定の処理を行うことで次元の圧縮を行うことが

†Department of Industrial Administration,Graduate School Of Science and Technology,Tokyo University of Science
‡Department of Industrial Administration,Faculty Of Science and Technology,Tokyo University of Science

できる． 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るものを同定対象の候補とする．
- STEP3
  候補のうち，文字数が 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．

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

4.1 データ

実験には BioCreAtIvEIII 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り，ACT とは PubMed の文献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る．これらのデータは専門家の手によって分類済みである．本研究ではトレーニングセットと開発セット，および BioCreAtIvEII 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，各文献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．

4.2 結果

実際の BioCreAtIvEIII 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 1 に示す．

表 1: 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
mcc	0.487	0.551
F-score	0.569	0.614

4.3 考察

本手法は既存の手法と比較して Sensitivity 以外は上回っていない．その理由の一つとして，タンパク質名の同定の精度が考えられる．ルールベースで行うことにより，当然データベースを用いるよりはタンパク質名同定の精度は下がる．そのため，判定が誤ってしまったということが考えられる．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文献に対しても，同様の性能を発揮することが期待でき，より汎用的な手法であるということが言える．

5 まとめ

本研究では，文献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を用いた手法を提案し BioCreAtIvE のデータセットに対し学習，分類を行った．その結果，既存の手法には及ばない点が多々あったが，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．また，ルールベースでの同定を行うことで特徴ベクトルの次元数の削減も実現した．

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(2011)

[2] 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(2001)

[3] K.Fukuda,T.Tsunoda,A.Tamura,T.Takagi:“Toward Information Extraction:Identifying protein names from biological papers”Pac Symp Biocomput,1998,707-18(1998)