トピックモデルを考慮したタンパク質間相互作用予測

東京理科大学大学院 理工学研究科 情報科学専攻
滝本研究室
6317632 宮崎 辰郎

平成 31 年 2 月 28 日

目 次

第1章　序論

1.1　背景

タンパク質には、単体で機能するものもあるが、そのほとんどが、他のタンパク質と相互作用することで機能を果たす。したがって、タンパク質の機能を解明する上で、タンパク質間相互作用 (*Protein Protein Interaction, PPI*) は必要不可欠なリソースである。 また、これらの相互作用に関する情報は、タンパク質間相互作用ネットワークを構築するために必要なリソースであり、生物学的プロセスの一般的な原理の理解を向上させることに役立てられている [1]。 したがって、タンパク質間相互作用に関する研究は、生物学の分野において非常に重要な研究に位置付けられている。 近年、研究の重要性と計算機の発達に伴い、計算機を用いたタンパク質間相互作用に関する研究が盛んに行われており、タンパク質間相互作用ネットワークの可視化 [2][3]、ネットワーク分析 [4]、未知のタンパク質間相互作用の予測問題 [5][6][7] などがある。 本稿では、これらのうち、未知のタンパク質間相互作用の予測問題について扱う。

通常タンパク質間相互作用は、実験によって検出され、実証されなければならない。 しかし、その実験は時間や金銭面において、非常にコストがかかってしまう。 その解決策の一つとして、既知のタンパク質間相互作用およびそれらで構築されるネットワークに基づいて未知のタンパク質相互作用を予測し、未知の相互作用が存在すると予測されたタンパク質間を中心に実験を行うことで、実験コストを大幅に削減することが期待できる。 そのため、未知のタンパク質間相互作用の予測問題に関する研究は重要視されている。

近年、機械学習やデータマイニングの領域では、ネットワークの構造で与えられるデータが増加しているので、その解析の重要性が高まり、ネットワーク構造から抽出される位相的情報やノード自身が持つ情報を活用する研究が活発に行われている。 データマイニングにおけるネットワーク解析をリンクマイニングと呼び、*Getoor* らは、リンクマイニングの問

題を以下のように分類している [8]。

Getoor らによるリンクマイニングの問題分類

1. ノードに関連する研究
   (a) ノードのランキング問題
   (b) ノードの分類問題
   (c) ノードのクラスタリング
   (d) ノードの識別
2. リンクに関連する研究
   (a) リンク予測問題
3. グラフに関連する研究
   (a) 部分グラフの検出
   (b) グラフの分類問題
   (c) グラフの生成モデル

これらのうち、本稿では、リンク予測問題を扱う。 リンク予測問題の詳細は、第 2.5 節で説明する。ノードをデータ、リンクをデータ間の関係として表現することで、様々な問題をリンク予測問題に適用することができる。 タンパク質間相互作用ネットワークでは、ノードはタンパク質、リンクはタンパク質間相互作用を表している。 したがって、既知のタンパク質間相互作用のネットワークの情報を用いて、未知のタンパク質間相互作用を予測する問題にリンク予測アルゴリズムを適用することができる。 これまでのリンク予測アルゴリズムを適用した研究では、タンパク質間相互作用ネットワークから得られる位相的情報が中心に扱われてきたが、近年、ノード自身が持つ情報が多く活用できるようになってきており、位相的情報に新規情報を加え、精度向上に貢献する研究が盛んに行われている。例えば、遺伝子発現ベースの手法 [9]、比較ゲノムベースの手法 [10]、コドンベースの手法 [11]、さらに、自然言語処理 (*natural language processing*) やテキストマイニングの技術を用いて抽出される意味的情報 [12] などが提案されている。これらに加え、本研究では自然言語処理などの領域で用いられるトピックモデル (*topic model*) を考慮した

タンパク質間に対する新規特徴量を提案する。 この新規特徴量を、位相的情報を含む既存の特徴量に加えることで、タンパク質間の表現力を高め、リンク予測問題の精度を向上できることを示す。

1.2 目的

本研究では、タンパク質間相互作用ネットワークから抽出する位相的情報や遺伝子オントロジー、アミノ酸配列から抽出されるタンパク質自身が持つ情報に加え、論文や学術誌からトピックモデルを生成することで得られる新しい特徴量を提案する。 本研究で提案する新規特徴量を用いた結果、未知のタンパク質間相互作用の予測精度の向上に貢献できることを示す。 予測精度の向上については、フリードマン検定及び多重比較を用いて、既存手法によって構成された特徴ベクトルと、提案手法であるトピックモデルを考慮した特徴ベクトルとの間に、統計的に有意差があることを示して議論する。 また、ランダムフォレストを用いて各特徴量の重要度を算出し、提案手法のタンパク質間相互作用予測における重要性について示す。

1.3 貢献

本研究の貢献は次の通りである。

- トピックモデルを考慮した特徴量を提案

本特徴量によって、従来の情報から得られた特徴量を用いたときに比べ、タンパク質間の表現力を高め、高い精度で未知のタンパク質間相互作用を予測することができる。

1.4 論文構成

第 2 章では、本稿で使用する諸概念について述べる。 第 3 章では既存手法による特徴量に加え、タンパク質間の表現力を高めることを目的としてトピックモデルを考慮した特徴量を提案する。 第 4 章では本研究で提案する特徴量の評価を行う。 第 5 章で関連研究について述べ、最後に第 6 章で本稿の結論を述べる。

第2章　準備

本章では、本稿で使用する諸概念について述べる。 まず、第 2.1 節でタンパク質、第 2.2 節でタンパク質間相互作用について説明し、第 2.3 節で分類問題、第 2.4 節で機械学習について説明する。 第 2.5 節でリンク予測について述べる。 第 2.6 節で仮説検定について説明し、最後に第 2.7 節でフリードマン検定とボンフェローニ法による多重比較検定について説明する。

2.1　タンパク質

タンパク質は、生物の体を構成している重要な成分の一つであり、炭水化物や脂質とともに三代栄養素と呼ばれる、20 種の L-アミノ酸が 50 個以上結合し、鎖状に多数連結してできた高分子化合物である。 結合している L-アミノ酸が 50 個未満の場合は、ペプチドと呼ばれる。 人間がタンパク質を摂取したあと、体内で分解され、アミノ酸になり体内に吸収される。

2.2　タンパク質-タンパク質間相互作用

タンパク質は、単独で機能するものもあるが、その多くは他のタンパク質と作用し合うことで機能する。 この、タンパク質分子間の相互作用のことをタンパク質-タンパク質間相互作用、もしくは簡潔にタンパク質間相互作用と呼ぶ。

2.3　分類問題

分類問題は 2 値分類、マルチクラス分類、マルチラベル分類の 3 つに大別される。 これは用意されるクラスの数、各事例が属するクラス数に

よって分けられる。

2.3.1 2値分類

2値分類は、各事例があるクラスに属するか属さないかの2種類で分類する方法である。 また、マルチクラス分類において、クラス数が2の場合と考えることができる。

例：数字 n を「奇数、偶数」のどちらかに分類

$$f(n) = \begin{cases} 0 & (n \text{が奇数}) \\ 1 & (n \text{が偶数}) \end{cases} \tag{2.1}$$

2.3.2 マルチクラス分類

マルチクラス分類は、クラス数が3以上の場合であり、各事例はどれか一つのクラスに分類される。

例：学生を所属学部で分類

クラス
　文学部、経済学部、法学部、教育学部、理工学部

としたとき、学生Aを経済学部、学生Bを理工学部と複数のクラスのうちのどれか一つに分類する。

2.3.3 マルチラベル分類

マルチラベル分類は、マルチクラス分類と同様に、クラス数が3以上である。 マルチクラス分類では、各事例はどれか一つのクラスに分類されるが、マルチラベル分類では、各事例が複数のクラスに同時に分類されてもよい。

例：*Wikipedia* の記事を分類
Wikipedia の「*Natural language processing*」の記事は「*Computational linguistics*」、「*Speech recognition*」、「*Natural language processing*」の3つのクラスに同時に分類されている。

2.4 機械学習の分類

機械学習は、教師あり学習 (*supervised learning*) と教師なし学習 (*unsupervised learning*) 、さらにその中間に位置する半教師あり学習 (*semi-supervised learning*) に大別できる。

2.4.1 教師あり学習

教師あり学習は、各入力データに対して出力 (正解ラベル) が付与されている場合に行う学習アプローチである。 正解ラベルは、教師データとも呼ばれる。 通常入力データは、実数値ベクトルとして表現される。 このベクトルは、各入力データの特徴を表すもので、特徴ベクトルと呼ばれる。 出力は、それぞれの事例の属するクラスを表すラベルや数値が与えられる。 出力がラベルのときを分類、数値のときを回帰と呼ぶ。 例えば天気予報は晴れ、曇り、雨などを予測するので、出力がラベルとなる。したがって、天気予報は、分類に相当する。 しかし、気温を予測する問題の場合、出力が数値となるので、回帰に相当する。

教師あり学習は、これらの入出力データを用いて、入力を出力に写像する関数や分布を求めることが目的である。したがって、$\boldsymbol{x} \in \boldsymbol{X}$ を入力、$\boldsymbol{y} \in \boldsymbol{Y}$ を出力とし、$|\boldsymbol{D}|$ 個の入出力ペアの集合 $\boldsymbol{D}$

$$\boldsymbol{D} = \{(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), (\boldsymbol{x}^{(2)}, \boldsymbol{y}^{(2)}), ..., (\boldsymbol{x}^{(|\boldsymbol{D}|)}, \boldsymbol{y}^{(|\boldsymbol{D}|)})\} \tag{2.2}$$

が与えられたとき、

$$f : \boldsymbol{X} \to \boldsymbol{Y} \tag{2.3}$$

となる f や、$\boldsymbol{x}$ が与えられたときの $\boldsymbol{y}$ の条件付き確率分布である、

$$p(\boldsymbol{y}|\boldsymbol{x}) \tag{2.4}$$

で表現される。 一般に f や p はパラメータを持ち、パラメータによって関数や分布が一意に決定される。 したがって学習は、これらのパラメータを決定する問題に帰着される。 以下で、教師あり学習の代表例として、サポートベクターマシンについて述べる。

サポートベクターマシン

サポートベクターマシン (*Support Vector Machine, SVM*) は教師あり学習を用いた線形二値分類である [13]。 また、カーネル法と組み合わせて用いることで、非線形の分類も可能であるが、ここでは、線形二値分類について説明する。 教師データに与える正解ラベルはそれぞれ、正クラス (*positive class*) 、負クラス (*negative class*) と呼ばれる。 また、正クラスに属する事例は正例 (*positive example*)、負クラスに属する事例は負例 (*negative example*) と呼ばれる。 事例の特徴ベクトル集合を $\{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(|\boldsymbol{D}|)}\}$、事例の正解ラベルを $\{y^{(1)}, y^{(2)}, ..., y^{(|\boldsymbol{D}|)}\}$ とする。このとき、正例のラベルは+1 であり、負例のラベルは-1 である。線形分類器の場合、分離平面の方向ベクトルを $\boldsymbol{\omega}$、切片を b とすると、

$$f(\boldsymbol{x}) = \boldsymbol{\omega} \cdot \boldsymbol{x} - b \tag{2.5}$$

と表される。事例 $\boldsymbol{x}$ を $f(\boldsymbol{x}) \geq 0$ ならば正クラス、$f(\boldsymbol{x}) < 0$ ならば負クラスに分類する。 このときのパラメータ $\boldsymbol{\omega}$ と b を求める方法の一つにマージン最大化がある。

例えば、図 2.1 のように事例が分布しているとする。 ここでは 2 次元空間を用いて説明するが、高次元でもかまわない。 この事例を分類する平面を分離平面と呼び、最適な分離平面を構築することが目的であるが、分離平面は無数に構築することができる。 例えば図 2.2 や図 2.3 のように構築することができる。

マージンとは分離面と分離面に最も近い事例を含む面との最短距離のことを言い、マージン最大化では、分離面に最も近い事例までのユーク

図 2.1: 事例の分布

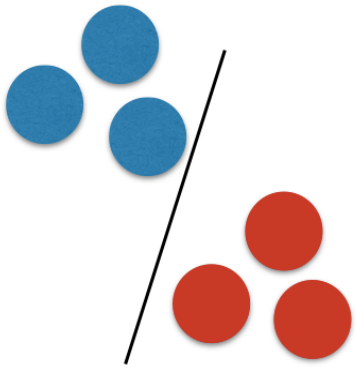


図 2.2: 分離平面：例 1

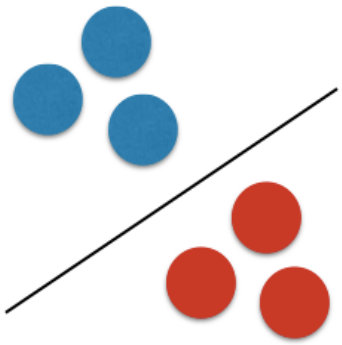


図 2.3: 分離平面：例 2

リッド距離が、最大となる分離面を決定する。 ここで、正例を 1 つ以上含む平面は、$\boldsymbol{\omega} \cdot \boldsymbol{x}_+ - b = +1$、負例を 1 つ以上含む平面は、$\boldsymbol{\omega} \cdot \boldsymbol{x}_- - b = -1$ で表され、これらの平面と分離平面との距離、つまりマージンは、$\frac{1}{||\boldsymbol{\omega}||}$ となる。 また $y^{(i)} = +1$ であるような事例については、$\boldsymbol{\omega} \cdot \boldsymbol{x^{(i)}} - b \geq 1$ であれば良く、$y^{(i)} = -1$ であるような事例については、$\boldsymbol{\omega} \cdot \boldsymbol{x^{(i)}} - b \leq 1$ であれば良い。 この 2 つの条件は、次のようにまとめて表すことができる。

$$y^{(i)}(\boldsymbol{\omega} \cdot \boldsymbol{x^{(i)}} - \boldsymbol{b}) \geq \mathbf{1} \tag{2.6}$$

以上から、これを制約とした次の最適化問題を解けばよい。

$$\begin{array}{ll} max. & \frac{1}{\boldsymbol{\omega}^2} \\ s.t. & y^{(i)}(\boldsymbol{\omega} \cdot \boldsymbol{x^{(i)}} - b) \geq 1 \\ & \forall i \end{array} \tag{2.7}$$

また、K 個のクラスに対して、あるクラスに入るか、他の $K-1$ 個のクラスのどれかに入るかの線形二値分類を解く分類器を K 個用いることでマルチクラスに拡張する *One-against-rest* が提案されている。

2.4.2 教師なし学習

教師なし学習では、入力データだけが与えられ、出力の正解は与えられない。 各データは、入力データの背後に存在する規則を見い出すために用いられ、どのような出力が望ましいかは、各学習アルゴリズムに依存する。 教師なし学習の適用対象の代表として、クラスタリングと頻出パターンマイニングの 2 つをあげられる。

クラスタリング

入力データ群から、類似したデータをグループ化する作業のことをクラスタリングと呼ぶ。 また、できあがったグループをクラスタと呼び、単純に最も類似しているもの同士をまとめる方法を凝集型クラスタリングと呼ぶ。

例：はじめは図 2.4 のように、すべての事例は互いに異なるクラスタに属しており、各クラスタはただ一つの事例だけを含む。 図 2.5、図 2.6、図 2.7 と進むに従って、類似しているもの同士をグループ化していき、少しずつ大きなクラスタが形成されていく。 また、この過程を図で表したものが図 2.8 の樹形図である。 クラスタリングの過程は樹形図の下から上に向かって進み、二つの線が交わっている箇所でクラスタ同士をまとめている。 交わる箇所の高さがまとめる順序を表している。 この樹形図をある高さで切ると、クラスタ集合が得られ、例えば上の方で切ると二つのクラスタが得られる。

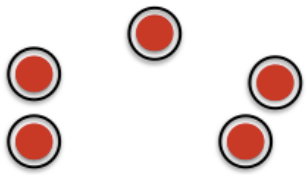


図 2.4: (a)

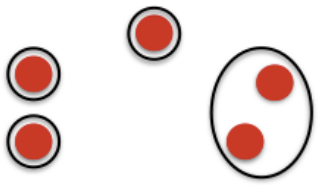


図 2.5: (b)

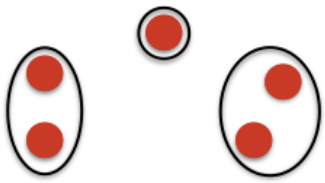


図 2.6: (c)

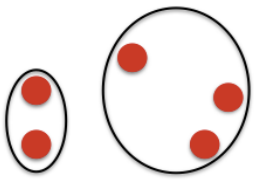


図 2.7: (d)

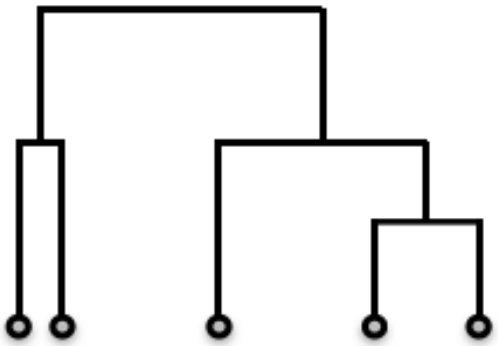


図 2.8: 樹形図

凝集型クラスタリング

はじめのうちは、類似している事例同士をまとめるが、クラスタリングが進むにつれて、類似するクラスタと事例、あるいはクラスタ同士をまとめることになる。 事例同士の場合は、事例の特徴ベクトルのユークリッド距離やコサイン類似度などを測り、最も類似しているクラスタ同士を見つければ良い。 一方、クラスタ同士の場合、つまり片方あるいは両方が一つのベクトルではなく複数のベクトルからなる場合、どのように計算したら良いかは自明ではない。 そこでクラスタ同士の類似度を測るための方法として、単連結法、完全連結法、重心法の 3 つの手法を紹介する。 扱っている問題によって、3 つの手法を選んで使う必要がある。各手法は、二つのクラスタの類似度を返す関数 sim に基づいて、次のように表される。

1. 単連結法

   二つのクラスタが与えられたとき、その中で最も近い事例対の類似度を、その二つのクラスタの類似度とする方法である。

$$sim(c_i, c_j) = \max_{\boldsymbol{x_k} \in c_i, \boldsymbol{x_l} \in c_j} sim(\boldsymbol{x_k}, \boldsymbol{x_l}) \tag{2.8}$$

式 (2.8) の値が最も大きなクラスタの対が、最も類似しているクラスタであるので、結合するクラスタは、

$$(c_i, c_j) = \arg\max_{c_i, c_j \in C} \max_{\boldsymbol{x_k} \in c_i, \boldsymbol{x_l} \in c_j} sim(\boldsymbol{x_k}, \boldsymbol{x_l}) \tag{2.9}$$

で与えられる c_i と c_j である。

2. 完全連結法

   二つのクラスタが与えられたとき、その中で類似度が最も低い事例同士の類似度を、その二つのクラスタの類似度とする方法を完全連結法と呼ぶ。

$$sim(c_i, c_j) = min_{\boldsymbol{x_k} \in c_i, \boldsymbol{x_l} \in c_j} sim(\boldsymbol{x_k}, \boldsymbol{x_l}) \tag{2.10}$$

3. 重心法

   各クラスタの代表ベクトルを、各クラスタが含む事例の重心ベクトルで表す。 二つのクラスタに対し、それらの代表ベクトルの類似度をこれらのクラスタ間の類似度とする方法を重心法と呼ぶ。

$$sim(c_i, c_j) = sim(\frac{1}{|c_i|} \sum_{\boldsymbol{x} \in c_i} \boldsymbol{x}, \frac{1}{|c_j|} \sum_{\boldsymbol{x} \in c_j} \boldsymbol{x}) \tag{2.11}$$

凝集型クラスタリングのアルゴリズムを次に示す。

Algorithm 1 凝集型クラスタリング

Input: 入力データ集合 $D = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, ..., \boldsymbol{x}^{(|D|)}\}$

1: $C = \{c_1, c_2, ..., c_{|D|}\}$
2: $c_1 = \{\boldsymbol{x}^{(1)}\}, c_2 = \{\boldsymbol{x}^{(2)}\}, ..., c_{|D|} = \{\boldsymbol{x}^{(|D|)}\}$
   # 1 つのクラスタに 1 つの事例を割り当てる
3: **while** $|C| \geq 2$ **do**
4: $(c_i, c_j) = \arg\max_{c_i, c_j \in C} sim(c_i, c_j)$
   #最も類似しているクラスタ対をみつける
5: $merge(c_i, c_j)$
   #見つかったクラスタ対をまとめる
6: **end while**

頻出パターンマイニング

入力データ群のうち、あるルールを満たすパターンで、出現頻度が高いパターンを列挙する手法を、頻出パターンマイニングと呼ぶ。 ルールの代表例として相関ルールがあげられる。 相関ルールとは、

$$X \Rightarrow Y \tag{2.12}$$

と記述され、ある事象 X が起こった下で、ある事象 Y が頻繁に発生する関係を表す。 このとき、X を前提部、Y を結論部と呼ぶ。 相関ルールを抽出するアルゴリズムの一つに、*Apriori*[14] が広く知られている。

2.4.3 半教師あり学習

半教師あり学習は、前述の教師あり学習と、教師なし学習の中間に位置する問題である。 これは、訓練データの中に、正解が与えられないようなデータが含まれていることを指す。 入力データは大量に存在するが、正解ラベルを付与するのに膨大なコストが必要である場合に有効である。以下に、教師なし学習の概要を示す。

教師なし学習の概要

1. ラベル付きデータを用いて分類器を構築する。
2. 1 で構築した分類器を用いてラベルなしデータを分類する。
3. 2 で分類したラベルなしデータのうち高い確率で分類できたものだけにラベルを付与する。

教師なし学習は、1 ～ 3 を繰り返し行うことで学習を行う。

2.5 リンク予測問題

機械学習やデータマイニングの分野では、ネットワーク構造の分析に関する研究が活発に行われている。 その一つがリンク予測である。 ネットワーク構造で表現されるデータが与えられ、その既知の位相情報から未知の位相情報を予測する問題である。 例えばソーシャルネットワークサービス (*SNS*) においては、与えられたネットワークの情報から、ユーザ同士の繋がりを予測する問題が例としてあげられる。

本節では、まず、第 2.5.1 項で、機械学習のアプローチをリンク予測問題に適用できることを述べ、次に、第 2.5.2 項で、機械学習を用いてリンク予測を行うときに用いる、ネットワーク構造から得られる位相的情報について説明する。 最後に、第 2.5.3 項で、それを特徴量とし 2 値分類に適応することを説明する。

2.5.1 機械学習アプローチを用いたリンク予測

リンク予測問題では、各リンクが存在するか否かを予測すれば良いので、入力をノード対の情報、出力を各ノード対におけるリンクの有無とする 2 値分類問題として考えることができる。 すなわち、各ノード対の特徴ベクトルを抽出し、機械学習アプローチを適用すれば良い。 したがって、各ノード対の特徴ベクトルを定義することが、機械学習アプローチを用いたリンク予測問題において重要なステップとなる。 各ノード対に用いることのできる特徴量は、ネットワークの位相的情報とノード自身が持つ情報の 2 つに大別される。

- ネットワークの位相的情報

  ノードに隣接している、またはその周辺に存在するノードのリンク構造から得られる情報である。 共通の知人を持つ人同士は、その人同士も知人である可能性が高いといった具合である。

- ノード自身が持つ情報

  例：ソーシャルネットワークサービスにおけるユーザの ID や年齢、趣味など タンパク質間相互作用ネットワークにおけるタンパク質の配列情報など

各リンクの有無を、2 値分類問題として扱うことで、機械学習のアプローチを、リンク予測に適用することが可能である。通常の機械学習アルゴリズムが、各ノードに対する特徴ベクトルを定義し、その特徴ベクトルを基にノードの性質などを予測しているのに対し、機械学習アプローチを用いたリンク予測では、各ノード対に対する特徴ベクトルを定義し、その特徴ベクトルを基にノード対の性質 (リンクの有無など) を予測している。

2.5.2 ネットワーク構造から得られる情報

リンク予測におけるネットワークとは、各データをノードとし、データ間の関係をリンクとして表現したものである。 ここでは、ネットワークに存在する任意の 2 つのノードの周辺ノードの情報を用いて定義される 7 つのノード間の類似尺度の定義を示す。

以下、i 番目のノードを x_i とし、x_i に隣接するノードの集合を $\Gamma(x_i)$ とする。

1. *common neighbors*

$$S^{CN}(x_i, x_j) \equiv |\Gamma(x_i) \cap \Gamma(x_j)| \tag{2.13}$$

common neighbors はリンク予測で最も使われる指標の一つである。*common neighbors* は x_i と x_j が共通の隣接ノードを多く持っているほど、x_i と x_j 間にはリンクが現れやすいことを表現している。 ソーシャルネットワークを例に挙げると、「共通の知人が多い 2 人は、互いに知人である可能性が高い」という具合である。

2. *Jaccard's coefficient* [15]

$$S^{JC}(x_i, x_j) \equiv \frac{|\Gamma(x_i) \cap \Gamma(x_j)|}{|\Gamma(x_i) \cup \Gamma(x_j)|} \tag{2.14}$$

Jaccard's coefficient は、*common neighbors* を正規化した式で得られる。すなわち、2 つの隣接ノード集合の和集合に対する共通の隣接ノード集合の割合を表現している。

3. *Adamic / Adar index* [16]

$$S^{AAI}(x_i, x_j) \equiv \sum_{z \in \Gamma(x_i) \cap \Gamma(x_j)} \frac{1}{\log |\Gamma(z)|} \tag{2.15}$$

Adamic / Adar index は、*Adamic* らが提案した、2 つのウェブページ間に対する初めての類似性尺度であり、*common neighbors* の重み付き和で表現され、隣接するノードが持つ隣接ノード数に応じて、重みが割り当てられる。 隣接ノードの少ないノードを共通して隣接ノードとして持つ場合、重みは大きくなり、隣接ノードの多いノー

ドを共通して隣接ノードとして持つ場合、重みは小さくなる。 ソーシャルネットワークを例に挙げると、「知人の少ない人の知人である二人は、その人同士も知人である可能性が高く、知人の多い人を共通の知人に持っていても、その人同士が知人であるかどうかを知るための情報には、不十分である」という具合である。

4. *Preferential attachment* [17]

$$S^{PA}(x_i, x_j) \equiv |\Gamma(x_i) \times \Gamma(x_j)| \tag{2.16}$$

Preferential attachnent は、「隣接ノードが多いノードほど、新たにリンクが張られやすい」ことを仮定している。 定義を見てわかる通り、*Preferential attachment* はそれぞれのノードの隣接するノードの集合の直積集合の要素数で与えられるが、結果として、2 つのノードが持つ隣接ノード数の掛け算と同値である。 したがって、共通の隣接ノードを持っている必要はない。

5. *Resource allocation* [18]

$$S^{RAI}(x_i, x_j) \equiv \sum_{z \in \Gamma(x_i) \cap \Gamma(x_j)} \frac{1}{|\Gamma(z)|} \tag{2.17}$$

Resource allocation は *Adamic / Adar index* に似ているが、*Adamic / Adar index* が重みを割り当てるときに対数の逆数を用いていたのに対し、*Resource allocation* は、隣接ノード集合の要素数の逆数で定義される。 これは、*Adamic / Adar index* に比べて、隣接するノード数が多いノードを共通の隣接するノードに持つ場合に、大きくペナルティを課していることを表している。

6. *Sorensen Dice coefficient* [19]

$$S^{SDC}(x_i, x_j) \equiv \frac{2|\Gamma(x_i) \cap \Gamma(x_j)|}{|\Gamma(x_i)| + |\Gamma(x_j)|} \tag{2.18}$$

Jaccard's coefficient では、2 つの隣接ノード集合の和集合を用いて *common neighbors* を正規化している。 しかし、2 つの隣接ノード集合の差集合の要素数が多いと、*Jaccard's coefficient* は、小さく見積もられてしまう。 *Sorensen Dice coefficient* は、この欠点を緩和するために、隣接ノード集合の積集合に重きをおき、差集合の要素数の影響を抑えることを目的として提案された。

7. *Overlap coefficient* [20]

$$S^{overlap}(x_i, x_j) \equiv \frac{|\Gamma(x_i) \cap \Gamma(x_j)|}{min\{|\Gamma(x_i)|, |\Gamma(x_j)|\}} \tag{2.19}$$

Jaccard's coefficient の欠点である、隣接ノード集合の差集合の影響を緩和するために、*Sorensen Dice coefficient* が提案されたが、それでもなお、差集合の要素数が膨大になった場合に、小さく見積もられてしまう。 *Overlap coefficient* では、差集合の要素数の影響を極限まで抑えることを目的として提案された。 *Jaccard's coefficient* や *Sorensen Dice coefficient* に比べ、隣接ノード集合の積集合の要素数に重きをおいた類似性尺度と言える。

2.5.3 ノード対に関する特徴量を用いた 2 値分類

ノード対 (x_i, x_j) に対する特徴量を、位相構造から得られる情報やノード自身が持つ情報などを組み合わせ、$\boldsymbol{v}^{(i,j)} = [v_1^{(i,j)}, v_2^{(i,j)}, \cdots, v_n^{(i,j)}]$ と表せたとする。 ここで、$v_k^{(i,j)}$ をノード x_i と x_j における k 番目のノード対に関する情報であり、n はノード対に関する情報の総数である。 例えば、第 2.5.2 項で説明した 7 つの位相的情報を用いるとき、$\boldsymbol{v}^{(i,j)} = [S^{CN}(x_i, x_j), S^{JC}(x_i, x_j), \cdots, S^{overlap}(x_i, x_j)]$、$n = 7$ となる。また、このノード対にリンクが存在することが既知であれば、ラベルを $y^{(i,j)} = 1$、リンクが存在しないことが既知であれば、0 とする。 これらの既知のノード対の情報を教師データとして与え、通常の機械学習アプローチ (ランダムフォレストやサポートベクターマシンなど) を適用することで、未知のノード対のリンクに関する知見を予測することができる。

2.6 仮説検定

本節では、まず第 2.6.1 項で仮説検定の考え方について述べ、そのあと第 2.6.2 項で帰無仮説と対立仮説について述べ、最後に、第 2.6.3 項で両側検定と片側検定検定について説明をする。

2.6.1 仮説検定の考え方

仮説検定 (*hypothesis testing*) とは、統計的仮説の有意性に関する検定である。 つまり、仮説のもとで期待されるものと、実際に観測された結

表 2.1: メンデルが行ったエンドウ豆についての実験データ

型	黄 · 丸	黄 · しわ	緑 · 丸	緑 · しわ	計
観測度数	315	101	108	32	556
理論比	9	3	3	1	(16)

果との間の違いが、偶然生じたものか否かを確率の基準で評価する。 例えば、1865 年にメンデルが行った、エンドウ豆の色や形状に関する遺伝形質を調査した実験で得られた実験データが、理論上の仮説に合致しているかどうかの検証があげられる。 この実験結果 (表 2.1) から、メンデルは各型の度数の比が 9:3:3:1 になると考えた。 しかし、厳密には 9:3:3:1 になっていないことは自明である。 ここで重要なことは、理論比からのずれが、誤差の範囲内なのかどうかという点である。 理論比からのずれが誤差の範囲内ではない場合、統計学では仮説は有意であるという。 ここで有意であるとは、偶然起こったとは判断しかねることを意味する。

もう一つ、コイン投げの例をあげて説明する。 例えば、コインを 10 回投げたときに 8 回表が出た場合、このコインに歪みがないと言えるかどうかという問題があげられる。 「歪みがない」という仮説のもとでは、$p = 1/2$、$N = 10$ であるので、$Bi(10, 1/2)$ の二項分布に従う。 もし仮説が正しい場合、表の回数を確率変数 X で表すと、二項分布の計算から、$P(X \geq 8) = 0.0537$ であるから、$X = 8$ は、「歪みがない」という仮説のもとでは、コインが 8 回 (以上) 出る確率は、約 5%である。 このことから、観測されるはずのない値であると主張することができる。 したがって、「歪みがない」という仮説が間違っているという結論を下す。 このことを、仮説を棄却 (*reject*) すると表現する。 しかし、約 5%でも確率が 0 ではないので、起こりうる確率であると主張することもできる。 そこで、仮説を棄却するときの基準、つまり、仮説が有意である確率の基準を定める必要がある。 この基準を有意水準 (*significance level*) といい α で表す。 有意水準 α を下回れば、起こり得ない確率とする。 今回のコインの例では、有意水準を $\alpha = 0.1$ にした場合、0.0537 は有意水準を下回るので、仮説は棄却される。 しかし、有意水準を $\alpha = 0.01$ にした場合、有意水準を上回るので帰無仮説は棄却されない。

コインの例で $P(X \geq 8)$ の確率を求めたが、この値を p 値と呼ぶ。p 値とは「仮説が正しいという仮定の下で、偏った統計検定量が得られる確率」を表している。 したがって、有意水準を p 値が下回ったときに仮説

を棄却し、統計学的には有意であると結論づけることができる。

2.6.2 帰無仮説と対立仮説

第 2.6.1 項では仮説検定の考え方について述べた。 例に挙げたコインを 10 回投げる試行において、コインに歪みがないという仮説は、有意水準 $\alpha = 0.1$ で棄却された。 このときにたてた仮説を、帰無仮説 (*null hypothesis*) とよび、H_0 で表す。 帰無仮説 H_0 のもとで計算された確率が有意水準を下回る場合には、帰無仮説 H_0 は正しくないと考える。

帰無仮説 H_0 が棄却されたときに採用する仮説を明示的に考えることがある。 それを対立仮説 (*alternative hypothesis*) とよび、H_1 で表す。

仮説検定では、標本から得られた統計量の値が、ある領域の値をとるときに、帰無仮説を棄却するのが一般的な手順であり、帰無仮説を棄却する範囲を棄却域 (*rejection region*) とよび、採択する領域を採択域 (*acceptance region*) とよぶ。

仮説検定の結果で積極的に主張できることは、帰無仮説 H_0 を棄却し対立仮説 H_1 が正しいと判断できるときだけである。 したがって、実験者が主張したいことを対立仮説にしておけば良い。 コインの例では、コインに歪みがあることを対立仮説とすれば良い。

まとめると、仮説検定とは、標本によって得られた情報を用いて、母集団に関する主張を採択するか棄却するかを判断することである。 H_0 のもと標本 $X = x$ が偶然生じたものなのか否かを、裾の確率を用いて測定し、その確率が有意水準 α を下回れば帰無仮説を棄却し、そうでない場合は帰無仮説を採択 (*accept*) する。

2.6.3 両側検定と片側検定

帰無仮説が差がない状態を表すのに対し、対立仮説は差がある状態を表す。 差の有無を考える場合と大小関係を考慮する場合に応じて、両側対立仮説と片側対立仮説がある。 両側対立仮説に対する検定を両側検定 (*two-sided test*)、片側対立仮説に対する検定を片側検定 (*one-sided test*) と呼ぶ。

コインの例では、コインに歪みがあることを主張したいので、帰無仮説 $H_0 : p = 1/2$、対立仮説 $H_1 : p \neq 1/2$ とたてて仮説検定を行う。 これは両側対立仮説の例であり、それに対応する検定は両側検定である。

表 2.2: 4 種類の肥料による収穫量

品種	肥料 1	肥料 2	肥料 3	肥料 4
品種 1	10	20	14	18
品種 2	5	30	14	9
品種 3	4	20	10	11

一方、試験勉強に効果があることを主張したいとする。 つまり、帰無仮説を H_0 :「試験勉強をする前とした後で試験の結果に差がない」とする。 このことを表現するため、試験の結果の変化量の母平均 μ に関して $H_0 : \mu = 0$ とすれば良い。 試験勉強に効果がある場合、μ は大きくなるので、対立仮説には $H_1 : \mu > 0$ を想定すれば良い。 これは、片側対立仮説の例であり、それに対応する検定は片側検定である。

2.7 フリードマン検定とボンフェローニ法による多重比較検定

本節では、第 2.7.1 項で、3 群以上の対応のあるデータに対する差の検定であるフリードマン検定について説明し、第 2.7.2 項でボンフェローニ法による多重比較検定について説明する。

2.7.1 フリードマン検定

フリードマン検定 (*Friedman Test*) とは、3 群以上の対応のあるデータについて、ノンパラメトリックで行う代表値の差の検定である。 帰無仮説 H_0 を「各処理対の母代表値に差はない」、対立仮説 H_1 を「各処理対の母代表値に差がある」とし、有意水準 α で両側検定を行う。

ここでは、4 種類の肥料間で収穫量に差があるかを検定することを例に挙げて説明する。 例で扱うデータは表 2.2 を参照されたい。

まずはじめに、帰無仮説と対立仮説を次のようにたてる。

- 帰無仮説 H_0 : 収穫量に差はない。
- 対立仮説 H_1 : 収穫量に差がある。

表 2.3: 4 種類の肥料に収穫量の順位

品種	肥料 1	肥料 2	肥料 3	肥料 4
品種 1	1	4	2	3
品種 2	1	4	3	2
品種 3	1	4	2	3
$R_{\cdot j}$	3	12	7	8
$R_{\cdot j}^2$	9	144	49	64

次に、各肥料ごとに収穫量 x_{ij} の小さい順に順位 R_{ij} をつける。 ただし、x_{ij} は j 番目の肥料を用いた場合の i 番目の品種の収穫量を表し、R_{ij} は x_{ij} の順位を表す。 同順位が存在する場合は、平均順位をそれぞれにつける。 一般に、対照群 (例の場合、品種) の数を n、処理群 (例の場合、肥料) の数を m とすると、$1 \leq i \leq n$、$1 \leq j \leq m$ であり、$1 \leq R_{ij} \leq m$ となる。

順位をつけ終えたら、各処理ごとに順位の和と順位の 2 乗和を計算する。

$$R_{\cdot j} = \sum_{i=1}^{n} R_{ij} \tag{2.20}$$

$$R_{\cdot j}^2 = \sum_{i=1}^{n} R_{ij}^2 \tag{2.21}$$

表 2.2 をもとに、各処理ごとの順位の和と順位の 2 乗和を計算した結果が表 2.3 である。

次に、式 (2.22) で定義された検定統計量 $\chi^2(m-1)$ を計算する。 検定統計量 $\chi^2(m-1)$ は自由度 $m-1$ の χ^2 分布に従う。

$$\chi^2(m-1) = \frac{12}{nm(m+1)} \sum_{j=1}^{m} R_{\cdot j}^2 - 3n(m+1) \tag{2.22}$$

肥料と収穫量の例では、$\chi^2(m-1) = 8.2$ であり自由度 3 の χ^2 分布に従う。

有意水準 α で検定を行う場合、$\chi^2(m-1) > \chi^2(m-1)_\alpha$ ならば帰無仮説 H_0 は有意水準 α で棄却され、$\chi^2(m-1) \leq \chi^2(m-1)_\alpha$ ならば帰無仮説 H_0 は棄却されない。 例の場合、自由度 3 の χ^2 分布において、$\chi^2(4-1)_{0.05} = 7.81$ なので、帰無仮説は棄却され、「肥料間に差がある」という結果を得る。

2.7.2　ボンフェローニ法による多重比較検定

本項では、まず、多重比較検定が必要な理由を述べ、そのあとに、本研究の実験で用いたボンフェローニ法について述べる。

多重比較検定が必要な理由

多重比較とは、3群以上の観測値において、どの群間に有意差があるのかを検定する方法である。 パラメトリック法では、母平均の差を、ノンパラメトリック法では、平均順位の差を比較する。 どの群間に有意差があるのかを検定するために群の各対で繰り返し検定を行う場合、多重性によって、定めた有意水準に比べて第1種の誤りの確率が大きくなってしまう。 例えば、A、B、Cの3群があったとして、有意水準αで検定を行うとき、各対(A, B)、(A, C)、(B, C)の差の検定で帰無仮説が採択される確率P_{AB}、P_{AC}、P_{BC}と棄却される確率P_{AB^c}、P_{AC^c}、P_{BC^c}はそれぞれ、

$$P_{AB} = P_{AC} = P_{BC} = 1 - \alpha \tag{2.23}$$

$$P_{AB^c} = P_{AC^c} = P_{BC^c} = \alpha \tag{2.24}$$

と表される。 第1種の誤りは、少なくとも一つの検定で帰無仮説が棄却されたときなので、その確率P_{er}は、

$$P_{er} = 1 - P_{AB} \cdot P_{AC} \cdot P_{BC} = 1 - (1 - \alpha)^3 \tag{2.25}$$

と計算される。 例えば、有意水準$\alpha = 0.05$で検定を行うとき、$P_{er} = 1 - (1 - 0.05)^3 = 0.142625$となるので、通常の第1種の誤りの確率の約3倍となってしまう。 これが、群の各対で検定を繰り返し行うことのできない理由である。

ボンフェローニ法

多重比較は、第1種の誤りの確率が大きくなることを防ぐ。 多重比較には統計量を用いた方法やp値や有意水準αを調整する方法があるが、ボ

ンフェローニ法 (*BonferroniCorrection*) は p 値や有意水準 α を調整する方法である。

具体的には、検定数 (検定する対の数) が N のとき、それぞれの p 値を $N \times p$ に調整して検定を行うか、検定の有意水準 α を α/N に変更する方法である。 例えば、10 対で検定を行う場合、各 p 値を $10 \times p$ にするか、全ての検定において $\alpha/10$ を有意水準として使う。

第3章　提案手法

本章では、第3.1節で文書中の単語のトピックを確率的に求める言語モデルである *LDA* について説明し、そのあとに、第3.2節で、提案手法である、トピックモデルを考慮したタンパク質を表現する新規情報について述べる。

3.1　Latent Dirichlet Allocation (LDA)

文書中の単語の潜在的トピックを確率的に推論するモデルを *LDA* と呼ぶ[21]。 *LDA* は、各単語は潜在的なトピックを持ち、同じトピックを持つ単語は同じ文書に出現しやすいことを仮定している。 その潜在的なトピックの分布の事前分布にディリクレ分布を仮定して生成する。 以下にディリクレ分布や多項分布の定義、*LDA* の生成過程、*LDA* のグラフィカルモデルを記す。

ディリクレ分布 (*Dirichlet distribution*)

$$Dir(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^{k}\alpha_i)}{\prod_{i=1}^{k}\Gamma(\alpha_i)}\prod_{i=1}^{k}\theta_i^{\alpha_i-1} \propto \prod_{i=1}^{k}\theta_i^{\alpha_i-1} \tag{3.1}$$

多項分布 (*Multinomial distribution*)

$$Multi(\boldsymbol{n}|\boldsymbol{\theta}) = \frac{\Gamma((\sum_{k=1}^{K}N_k)+1)}{\prod_{k=1}^{K}\Gamma(N_k+1)}\prod_{k=1}^{K}\theta_k^{N_k} \tag{3.2}$$

LDA の生成過程

トピックの確率分布を選択 $\boldsymbol{\theta}_m \sim Dir(\boldsymbol{\theta}_m|\boldsymbol{\alpha}) \quad (m = 1, ..., M)$
トピックごとに単語出現確率分布を生成 $\boldsymbol{\phi_k} \sim Dir(\boldsymbol{\phi}_k|\boldsymbol{\beta}) \quad (k = 1, ..., K)$
単語の潜在的トピックを生成 $z_{m,n} \sim Multi(z_{m,n}|\boldsymbol{\theta}_m) \quad (n = 1, ..., N_m)$
トピックに応じた確率で単語を生成 $w_{m,n} \sim Multi(w_{m,n}|\boldsymbol{\theta}_{z_{m,n}}) \quad (n = 1, ..., N_m)$

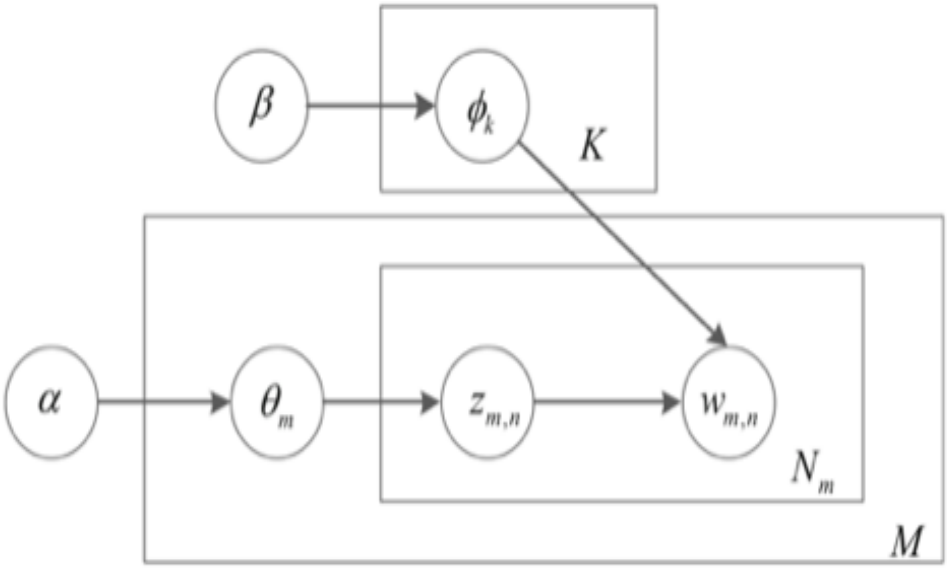


図 3.1: *LDA* のグラフィカルモデル

グラフィカルモデルにおける各変数の意味は次のようになる。

- M：文書数
- N_m：文書 m に含まれる単語数
- k：トピック数
- $w_{m,n}$：文書 m、単語 n の語彙インデックス
- $z_{m,n} \in [1, \cdots, k]$：文書 m における単語 n のトピック番号
- $\boldsymbol{\theta}_m$：文書 m におけるトピックの出現確率
- $\boldsymbol{\phi}_k$：トピック k における語彙の出現確率
- $\boldsymbol{\alpha}$：トピックの出現確率の偏りを表すパラメータ
- $\boldsymbol{\beta}$：語彙の出現確率の偏りを表すパラメータ

図 3.1 のグラフィカルモデルから、$\boldsymbol{\theta} \to \boldsymbol{z} \to \boldsymbol{w}$ の順で生成されることが分かる。 また、$\boldsymbol{\theta}$ は M 回、$\boldsymbol{z}$ は N 回生成される。 さらに、単語分布 $p(\boldsymbol{w}|\boldsymbol{z})$、トピックの選択確率 $p(\boldsymbol{z}|\boldsymbol{\theta})$、トピック分布 $p(\boldsymbol{\theta}|\alpha)$ を用いると、$\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}$ の同時確率は、

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}) = p(\boldsymbol{w}|\boldsymbol{z})p(\boldsymbol{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\alpha) \tag{3.3}$$

と表される。 したがって、文書 $\boldsymbol{w}_m = \{w_{m,1}, \cdots, w_{m,N_m}\}$、文書全体 $\boldsymbol{w} = \{\boldsymbol{w}_1, \cdots, \boldsymbol{w}_M\}$ についてはそれぞれ、

$$p(\boldsymbol{w}_m, \boldsymbol{z}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|\alpha) \prod_n p(w_{m,n}|z_{m,n})p(z_{m,n}|\boldsymbol{\theta}) \tag{3.4}$$

$$\begin{aligned} p(\boldsymbol{w}) =& \int \sum_{\boldsymbol{z}} p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}) d\boldsymbol{\theta} \\ =& \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int (\prod_k \theta_k^{\alpha_k - 1}) \prod_n \sum_k p(w_n|k)\theta_k d\boldsymbol{\theta} \end{aligned} \tag{3.5}$$

と計算できる。 式 (3.5) の積分は変分ベイズ法などを用いて推定する。

3.2 提案手法

本研究では、タンパク質間相互作用ネットワークから抽出される位相的情報や、遺伝子オントロジー、アミノ酸配列から抽出されるタンパク

質自身が持つ情報といった既存の特徴量に加え、論文や学術誌からトピックモデルを生成することで得られる新しい特徴量を提案する。本手法は、類似するタンパク質は、そのタンパク質について書かれている学術誌同士も、類似するトピックを含んでいることを仮定している。

まず、各タンパク質について記載されている学術誌を収集する。そのあと、収集した全学術誌をコーパスとし、各タンパク質のトピック分布を生成する。具体的には、各タンパク質について論述している学術誌の潜在的トピックを *LDA* を用いて確率的に推論することで、各タンパク質に関するトピックの推定を実現する。*LDA* における変数 $M, N_m, w_{m,n}, z_{m,n}, \boldsymbol{\theta}_m$ を本手法では、以下のように割り当てる。

- M' : タンパク質数
- N'_m : タンパク質 m について書かれている学術誌に含まれる単語数
- $w'_{m,n}$: タンパク質 m について書かれている学術誌に含まれる単語 n の語彙インデックス
- $z'_{m,n} \in [1, \cdots, k]$: タンパク質 m について書かれている学術誌における単語 n のトピック番号
- $\boldsymbol{\theta}'_m$: タンパク質 m について書かれている学術誌におけるトピックの出現確率

さらに、単語分布、トピックの選択確率、トピック分布はそれぞれ、$p(\boldsymbol{w}'|\boldsymbol{z}')$、$p(\boldsymbol{z}'|\boldsymbol{\theta}')$、$p(\boldsymbol{\theta}'|\alpha)$ となる。したがって、$\boldsymbol{w}'$、$\boldsymbol{z}'$、$\boldsymbol{\theta}'$ の同時確率は、

$$p(\boldsymbol{w}', \boldsymbol{z}', \boldsymbol{\theta}') = p(\boldsymbol{w}'|\boldsymbol{z}')p(\boldsymbol{z}'|\boldsymbol{\theta}')p(\boldsymbol{\theta}'|\alpha) \tag{3.6}$$

と表される。また、タンパク質 m について書かれている学術誌に出現する単語集合 $\boldsymbol{w}'_m = \{w'_{m,1}, \cdots, w'_{m,N'_m}\}$ と、収集した学術誌 $\boldsymbol{w}' = \{\boldsymbol{w}'_1, \cdots, \boldsymbol{w}'_{M'}\}$ についてはそれぞれ、

$$p(\boldsymbol{w}'_m, \boldsymbol{z}', \boldsymbol{\theta}') = p(\boldsymbol{\theta}'|\alpha) \prod_n p(w'_{m,n}|z'_{m,n})p(z'_{m,n}|\boldsymbol{\theta}') \tag{3.7}$$

$$
\begin{aligned}
p(\boldsymbol{w}') =& \int \sum_{\boldsymbol{z}'} p(\boldsymbol{w}', \boldsymbol{z}', \boldsymbol{\theta}') d\boldsymbol{\theta}' \\
=& \frac{\Gamma(\sum_k \alpha_k)}{\prod_k \Gamma(\alpha_k)} \int (\prod_k \theta_k'^{\alpha_k - 1}) \prod_n \sum_k p(w_n'|k) \theta_k' d\boldsymbol{\theta}'
\end{aligned}
\tag{3.8}
$$

と計算できる。 式 (3.8) も式 (3.5) と同様に、変分ベイズ法などを用いて推定すれば良い。

生成されたトピックモデルが各タンパク質を意味的観点から表現していると仮定し、全タンパク質間についてトピック分布を用いてコサイン類似度を測ることで、タンパク質間の新しい類似度を測ることを可能にする。 この類似度を新しい特徴量として加えることで、タンパク質間の表現力を高め、精度向上に貢献できると考える。図 3.2 に、本手法のフローチャートを示す。

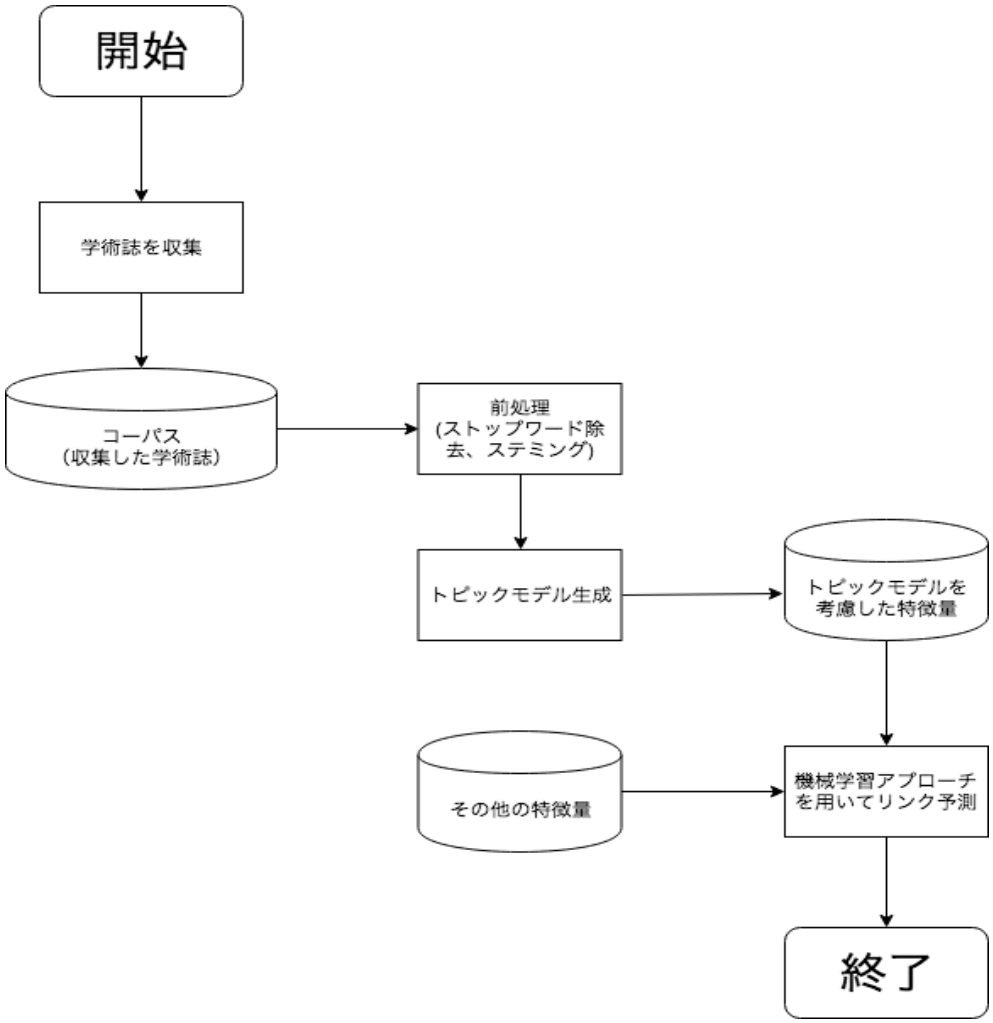


図 3.2: 提案手法のフローチャート

第4章　評価

本章では評価結果を示す。 まず、第4.1節で実験に用いるデータセットについて述べる。 第4.2節で評価指標について説明し、第4.3節で実験で用いた既存手法について説明する。 第4.4節で評価方法について説明したあと、第4.5節で実験結果を示す。 次に、第4.6節でフリードマン検定とボンフェローニ法による多重比較検定を用いた評価について示し、第4.7節で特徴量の重要度についての評価を示す。これらの結果を踏まえて第4.8節で提案手法について議論し、最後に、第4.9節で今後の課題について述べる。

4.1　データセット

本研究では、*STRING*[3] と呼ばれるタンパク質間相互作用に関するデータベースから、ランダムに3000個のタンパク質の情報を抽出して実験を行なった。 ただし、今回扱ったタンパク質は、ホモ・サピエンスに関するものである。 まず、*STRING*から、各タンパク質間の相互作用に関する情報を抽出した。 *STRING*から抽出される情報は、各タンパク質間の相互作用が存在するという情報に限るので、未知の相互作用を表現するために、既知の相互作用の10%をランダムに取り除いた。 この10%を未知の相互作用とし、予測することを実験の目的とする。

既知の相互作用を取り除いたあとのタンパク質間相互作用ネットワークを用いて、第2.5.2項で説明した各タンパク質間の位相的構造から得られる情報や、第4.3.1項、第4.3.2項で説明するタンパク質自身が持つ情報、さらに、提案手法であるトピックモデルを考慮した情報を用いて特徴ベクトルを生成し、7つの機械学習アプローチ (ランダムフォレスト、k-最近傍法、決定木、サポートベクターマシン、ナイーブベイズ分類器、ロジスティック回帰、ニューラルネットワーク) で精度評価を行なった。 全タンパク質対 (4498500対) のうち、4048650対 (正例：119190対、負例：3929460対) を教師データ、449850対 (正例：13243対、負例：436607

表 4.1: タンパク質間相互作用ネットワークの情報

ノード数 $\|V\|$	リンク数 $\|E\|$	平均次数 c	直径 $\|D\|$	平均パス長 $\|L\|$	クラスタ係数 C
3000	132433	88.29	6	2.46	0.30

表 4.2: クラス C

	C に属する	C に属さない
C であると予測	a	b
C でないと予測	c	d

対) をテストデータとして用いた。 今回用いたデータ中のタンパク質が構成するタンパク質間相互作用ネットワークの情報を表 4.1 に示す。

4.2 評価指標

実験の評価には、適合率 (*Precision*) と再現率 (*Recall*) とそれらの調和平均で与えられる f値を用いる。適合率は、正と予測したデータのうち、実際に正であるものの割合、再現率は実際に正であるもののうち、正であると予測されたものの割合を示す。例えば、クラス C に対する結果が表 4.2 のように得られているとする。ここで C に属すると予測したもののうち、C に属しているものの数を a、C に属すると予測したもののうち、C に属していないものの数を b、C に属さないと予測したもののうち、C に属しているものの数を c、C に属していないもののうち、C に属していないものの数を d で表す。

この分割結果をもとに適合率、再現率、f 値は次のように定義される。

$$Precision = \frac{a}{a+b} \tag{4.1}$$

$$Recall = \frac{a}{a+c} \tag{4.2}$$

$$f\text{値} = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4.3}$$

4.3 実験で用いた既存手法

本節では、実験で用いた既存の特徴量について説明をする。 第 4.3.1 項でタンパク質のアミノ酸配列におけるアライメントを用いた特徴量について説明し、第 4.3.2 項で遺伝子オントロジーの情報を考慮した特徴量について説明する。

4.3.1 タンパク質のアミノ酸配列におけるアライメントベース

バイオインフォマティクスにおける基本原理として、アミノ酸配列が似ていれば機能も似ている傾向がある。 その基本原理に基づいて用いられる技術の一つに最適アライメントがある。 最適アライメントとは、可能なアライメントの中でスコアが最大なものを求める手法である。 タンパク質が持つアミノ酸配列を 1 つの文字列として考えることで、タンパク質の類似性を定量的に算出することを可能にする。 可能なアライメントの数は指数個存在するので、全てのアライメントに対するスコアを算出するためには指数時間必要である。 そこで動的計画法を用いて最大スコアを求めるアルゴリズムが提案された。

動的計画法を用いたアルゴリズムについて説明する。$F(n, m)$ を配列 $x_1, x_2, \cdots, x_n$ と $y_1, y_2, \cdots, y_m$ の最適アライメントのスコアとする。 まず、式 (4.4) で初期化する。 gap はギャップペナルティと呼ばれ、定数で与えられる。

$$\begin{cases} gap = const. \\ F(0, j) = -j \times gap \quad (j = 1, \cdots, m) \\ F(i, 0) = -i \times gap \quad (i = 1, \cdots, n) \end{cases} \tag{4.4}$$

漸化式は式 (4.5) で与えられる。 $s(x_i, y_j)$ とは置換スコア (行列) と呼ばれ、各文字間の類似性を定量化している。 アミノ酸配列に対するスコア行列は、進化の過程における相対的な置換のしやすさを反映している。代表的な置換スコア行列として、*BLOSUM* 行列 [22] や *PAM* 行列 [23] が用いられているが、本研究では *BLOSUM* 行列 (図 4.1) を用いる。

	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Thr	Trp	Tyr	Val
Ala	4																			
Arg	−1	5																		
Asn	−2	0	6																	
Asp	−2	−2	1	6																
Cys	0	−3	−3	−3	9															
Gln	−1	1	0	0	−3	5														
Glu	−1	0	0	2	−4	2	5													
Gly	0	−2	0	−1	−3	−2	−2	6												
His	−2	0	1	−1	−3	0	0	−2	8											
Ile	−1	−3	−3	−3	−1	−3	−3	−4	−3	4										
Leu	−1	−2	−3	−4	−1	−2	−3	−4	−3	2	4									
Lys	−1	2	0	−1	−3	1	1	−2	−1	−3	−2	5								
Met	−1	−1	−2	−3	−1	0	−2	−3	−2	1	2	−1	5							
Phe	−2	−3	−3	−3	−2	−3	−3	−3	−1	0	0	−3	0	6						
Pro	−1	−2	−2	−1	−3	−1	−1	−2	−2	−3	−3	−1	−2	−4	7					
Ser	1	−1	1	0	−1	0	0	0	−1	−2	−2	0	−1	−2	−1	4				
Thr	0	−1	0	−1	−1	−1	−1	−2	−2	−1	−1	−1	−1	−2	−1	1	5			
Trp	−3	−3	−4	−4	−2	−2	−3	−2	−2	−3	−2	−3	−1	1	−4	−3	−2	11		
Tyr	−2	−2	−2	−3	−2	−1	−2	−3	2	−1	−1	−2	−1	3	−3	−2	−2	2	7	
Val	0	−3	−3	−3	−1	−2	−2	−3	−3	3	1	−2	1	−1	−2	−2	0	−3	−1	4

図 4.1: BLOSUM62

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + s(i,j) \\ F(i-1,j) - gap \\ F(i,j-1) - gap \end{cases} \tag{4.5}$$

4.3.2 遺伝子オントロジーベース

本項では最初に、遺伝子オントロジーについて説明し、次に、実験で用いる、*tf-idf* (*Term Frequency-Inverse Document Frecuency*) の考えを用いた類似尺度について説明する。 その他の遺伝子オントロジーに基づく類似尺度については 5.7 節で説明する。

遺伝子オントロジー

遺伝子オントロジー (*Gene Ontology*) とは、遺伝子の生物学的プロセス、細胞の構成要素、分子機能を基に遺伝子に付与されるアノテーションのことである [24]。 遺伝子オントロジーは階層構造を持っており、最上位の階層は以下の 3 つで分類されている。

- *Biological Process* (生物学的プロセス)

- *Cellular Component* (細胞の構成要素)
- *Molecular Function* (分子機能)

遺伝子オントロジーで定義された用語を *GO term* と呼び、それぞれに ID が割り当てられている (例： *GO : 0004888*、*GO : 0005515*)。遺伝子オントロジーは有向非巡回グラフで表現され、ノードが各 *GO term* を、リンクがそれらの *GO term* 同士に関係があることを示している [25]。つまり、各 *GO term* は一つ以上の親に相当する *GO term* と「$is-a$」、「$part-of$」関係をもつ。

遺伝子オントロジーを考慮した特徴量

Huang らは、多くの遺伝子に付与されている *GO term* はそれほど重要ではなく、あまり付与されていない *GO term* の情報を重要視するために、*tf-idf* の考えを用いた類似尺度の提案をした [26]。

Huang らは有向非巡回グラフ内の全ての用語の寄与を示すために、各用語 t の先祖ノード集合を以下のように定義した。

$$EV(t) = \{t'; t' \in DAG(t)\} \tag{4.6}$$

ただし、*DAG(t)* は t の自分自身と先祖ノードを含む *GO term* 集合である。

図 4.2 に、*cell-cell adhesion* が形成する有向非巡回グラフの例を示す。図 4.2 の場合、「*cell-cell adhesion (GO 0016337)*」の先祖ノードは「*cell adhesion (GO : 0007155)*」、「*celluar process (GO : 0009987)*」、「*biological adhesion (GO : 0022610)*」、「*biological process (GO : 0008150)*」である。したがって (*GO:0007155*) の先祖ノード集合は、$EV(GO : 0007155) = \{GO : 0007155, GO : 0009987, GO : 0022610, GO : 008150\}$ となる。

次に、ある遺伝子 G にアノテーションされている *GO term* が複数の先祖ノード集合に出現する可能性があるので、その出現頻度の情報を含む集合を次のように定義した。

$$EA(G) = \{(t, freq); t \in G_a\} \tag{4.7}$$

ただし、G_a は遺伝子 G にアノテーションされている *GO term* とそれらの先祖ノード集合、*freq* は G_a 内における *GO term* t の出現頻度を表す。*(GO : 0009987)* と *(GO : 002610)* がアノテーションされている遺伝子 G を考えた場合、図 4.2 では、*(GO : 0008150)* は *(GO : 0009087)* と *(GO :*

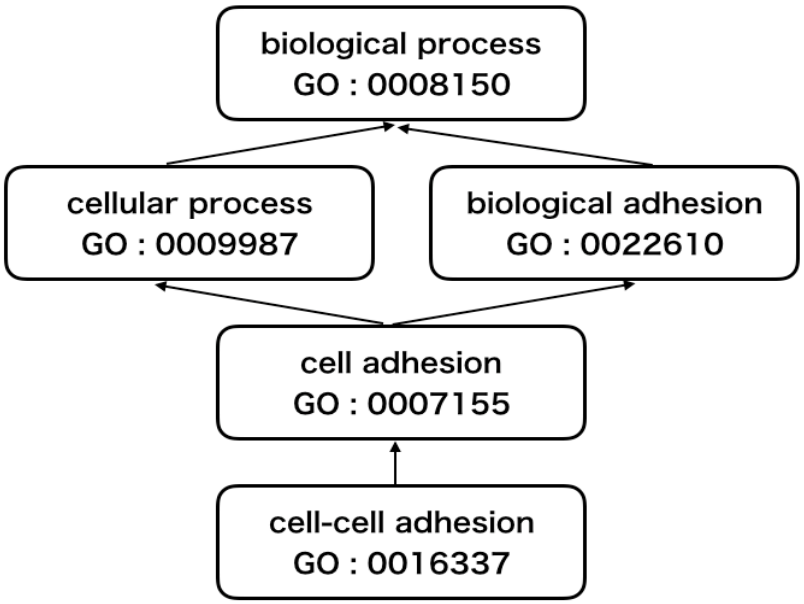


図 4.2: *cell-cell adhesion (GO:0016337)* が形成する有向非巡回グラフ

0022610) の先祖ノード集合で出現する。 したがって、$EA(G) = \{(GO : 0009987, 1), (GO : 0022610, 1), (GO : 0008150, 2)\}$ となる。

最後に、これらの定義された集合 (式 (4.6) や式 (4.7)) に基づいて、遺伝子 G にアノテーションされている *GO term* t に対する重みを *tf-idf* の考えを考慮し、以下のように計算する。

$$w_{t,G} = TF_{t,G} \cdot IDF_t = \frac{f_{t,G}}{max_z f_{t,G}} \ln(\frac{N}{n_t} + 0.01) \tag{4.8}$$

ただし、f_t は、式 (4.7) で得られた $EA(G)$ における $freq$、z は遺伝子 G にアノテーションされている *GO term* の数、N はデータセット中の全遺伝子数、n_t は *GO term* t が出現する遺伝子の数である。

遺伝子 G_i における *GO term* t_i の重みを w_{t_i,G_i} としたとき、$G_i = [w_{t_1,G_i}, w_{t_2,G_i}, \cdots, w_{t_n,G_i}]$ としてベクトル表現することが可能である。 ただし、ここで n はデータセット中に存在する *GO term* の総数を表す。

各遺伝子のベクトルを用いて式 (4.9) で与えられるコサイン類似度を計算することで、遺伝子の類似度を計算する。

$$sim(G_i, G_j) = \frac{\sum_{k=1}^{n} w_{t_k,G_i} w_{t_k,G_j}}{\sqrt{\sum_{k=1}^{n} w_{t_k,G_i}^2} \sqrt{\sum_{k=1}^{n} w_{t_k,G_j}^2}} \tag{4.9}$$

コサイン類似度は、0 から 1 の間の値をとり、1 に近いほどそれらの遺伝子が類似していることを表す。

4.4 評価方法

トピックモデルを考慮した新規特徴量を既存の特徴量 (位相的構造から得られる情報や、タンパク質自身が持つ情報) に組み込むことで、精度向上に貢献することを確認するために、既存の特徴量とのあらゆる組み合わせによって構成される特徴ベクトルを用いて、ランダムフォレスト、k-最近傍法、決定木、サポートベクターマシン、ナイーブベイズ分類器、ロジスティック回帰、ニューラルネットワークの 7 つの手法で精度評価を行った。 その際、10-分割交差検定を行なった。 その後、フリードマン検定および多重検定を用いて、トピックモデルを考慮した特徴量を含む特徴ベクトルと、その他の特徴ベクトルとの間に有意差があることを確認し、どの特徴ベクトル間に有意差があるかを示す。 さらに、ランダムフォレストを用いて、各特徴量の重要度を算出し、リンク予測における提案手法の重要度に関する評価を行なった。

今回実験で用いた特徴ベクトルを表 4.3 に示す。 本実験でタンパク質のトピックモデルを生成するために収集した文献は、$PubMed$ と呼ばれる医学や生物学の分野に関する学術誌への検索エンジンに掲載されている学術誌のアブストラクトをコーパスとして用いた。 そのあと、前処理として、ストップワードの除去とステミングを行い、LDA を用いてトピックモデルを生成した。

表 4.3: 実験に用いた特徴ベクトル

$\boldsymbol{x}_1$	*topology*
$\boldsymbol{x}_2$	*topology* + *alignment*
$\boldsymbol{x}_3$	*topology* + *GO*
$\boldsymbol{x}_4$	*topology* + *topic model*
$\boldsymbol{x}_5$	*topology* + *alignment* + *topic model*
$\boldsymbol{x}_6$	*topology* + *GO* + *topic model*
$\boldsymbol{x}_7$	*topology* + *alignment* + *GO*
$\boldsymbol{x}_8$	*topology* + *alignment* + *GO* + *topic model*

topology : 位相的構造から得られる情報、*alignment* : アミノ酸配列のアライメント情報、*GO* : 遺伝子オントロジーに基づく情報、*topic model* : トピックモデル

4.5 実験結果

表 4.4 に、各特徴ベクトルと、各分類器に対する実験結果を示す。 ただし、表中の *RF* はランダムフォレスト、*k-NN* は *k*-最近傍法、*DT* は決定木、*SVM* はサポートベクターマシン、*NB* はナイーブベイズ分類器、*LR* はロジスティック回帰、*NN* はニューラルネットワークを示している。

4.6 フリードマン検定と多重比較による評価

本節では、まず第 4.6.1 項で特徴ベクトル間に有意差があることをフリードマン検定を適用して確認し、そのあと第 4.6.2 項でどの特徴ベクトル間に有意差が生じているのかをボンフェローニ法による多重比較検定を用いて確認する。

4.6.1 フリードマン検定

特徴ベクトル間における精度の差の検定を、フリードマン検定を用いて行なった。 精度は、適合率と再現率の調和平均である *f* 値を用いた。 フリードマン検定は *IBM SPSS Statics* を用いて行なった。 その結果、特徴量ベクトル間に有意差が認められた ($\chi^2(7) = 44.284, p < 0.01$)。

表 4.4: 実験結果

特徴ベクトル	分類器	適合率	再現率	f値
$\boldsymbol{x}_1$	*RF*	0.69	0.56	0.62
	k-NN	0.60	0.54	0.57
	DT	0.70	0.54	0.61
	SVM	0.76	0.54	0.63
	NB	0.59	0.50	0.54
	LR	0.60	0.51	0.55
	NN	0.69	0.60	0.64
$\boldsymbol{x}_2$	*RF*	0.70	0.59	0.64
	k-NN	0.63	0.57	0.60
	DT	0.69	0.60	0.64
	SVM	0.68	0.62	0.65
	NB	0.60	0.53	0.56
	LR	0.60	0.56	0.58
	NN	0.71	0.60	0.65
$\boldsymbol{x}_3$	*RF*	0.71	0.62	0.66
	k-NN	0.69	0.56	0.62
	DT	0.71	0.60	0.65
	SVM	0.68	0.64	0.66
	NB	0.61	0.53	0.57
	LR	0.64	0.55	0.59
	NN	0.74	0.65	0.69
$\boldsymbol{x}_4$	*RF*	0.70	0.64	0.67
	k-NN	0.65	0.59	0.62
	DT	0.71	0.60	0.65
	SVM	0.74	0.64	0.69
	NB	0.60	0.56	0.58
	LR	0.63	0.57	0.60
	NN	0.72	0.66	0.69

特徴ベクトル	分類器	適合率	再現率	f値
$\boldsymbol{x}_5$	RF	0.75	0.59	0.66
	k-NN	0.71	0.55	0.62
	DT	0.71	0.60	0.65
	SVM	0.74	0.63	0.68
	NB	0.63	0.52	0.57
	LR	0.63	0.55	0.59
	NN	0.71	0.65	0.68
$\boldsymbol{x}_6$	RF	0.73	0.62	0.67
	k-NN	0.66	0.58	0.62
	DT	0.77	0.56	0.65
	SVM	0.74	0.65	0.69
	NB	0.63	0.54	0.58
	LR	0.65	0.56	0.60
	NN	0.74	0.65	0.69
$\boldsymbol{x}_7$	RF	0.71	0.60	0.65
	k-NN	0.66	0.57	0.61
	DT	0.76	0.54	0.63
	SVM	0.69	0.65	0.67
	NB	0.61	0.52	0.56
	LR	0.62	0.53	0.57
	NN	0.69	0.60	0.64
$\boldsymbol{x}_8$	RF	0.71	0.62	0.66
	k-NN	0.69	0.56	0.62
	DT	0.72	0.59	0.65
	SVM	0.73	0.64	0.68
	NB	0.60	0.54	0.57
	LR	0.62	0.56	0.59
	NN	0.70	0.66	0.68

表 4.5: 各特徴ベクトル対の調整済み有意確率と検定結果

特徴ベクトル	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\boldsymbol{x}_3$	$\boldsymbol{x}_4$	$\boldsymbol{x}_5$	$\boldsymbol{x}_6$	$\boldsymbol{x}_7$	$\boldsymbol{x}_8$
$\boldsymbol{x}_1$		0.395	**0.036**	**< 0.001**	**0.036**	**< 0.001**	1.000	**0.036**
$\boldsymbol{x}_2$	n.s.		0.934	**0.016**	0.934	**0.016**	1.000	0.934
$\boldsymbol{x}_3$	*	n.s.		1.000	1.000	1.000	1.000	1.000
$\boldsymbol{x}_4$	**	*	n.s.		1.000	1.000	**0.020**	1.000
$\boldsymbol{x}_5$	*	n.s.	n.s.	n.s.		1.000	1.000	1.000
$\boldsymbol{x}_6$	**	*	n.s.	n.s.	n.s.		**0.020**	1.000
$\boldsymbol{x}_7$	n.s.	n.s.	n.s.	*	n.s.	*		1.000
$\boldsymbol{x}_8$	*	n.s.	n.s.	n.s.	n.s.	n.s.	n.s.	

n.s. : 非有意、 * : 有意 ($p < 0.05$)、 ** : 有意 ($p < 0.01$)

4.6.2 多重比較

第 4.6.1 項で特徴ベクトル間に差があるという結果が得られたので、どの特徴ベクトル間に有意差があるのかを、多重比較で確認した。 多重比較は、*IBM SPSS Statics* を用いて行なった。 すべての対比較を行う際に有意水準を調整するために、ボンフェローニ法による多重比較検定を行なった。 *IBM SPSS Statics* を用いて求めた各特徴ベクトル対の調整済み有意確率と検定結果を表 4.5 に示す。 上三角の部分が、各特徴ベクトル対における調整済みの有意確率であり、下三角の部分が、検定結果である。 また、各特徴ベクトルの精度に関する箱ひげ図を図 4.3 に示す。

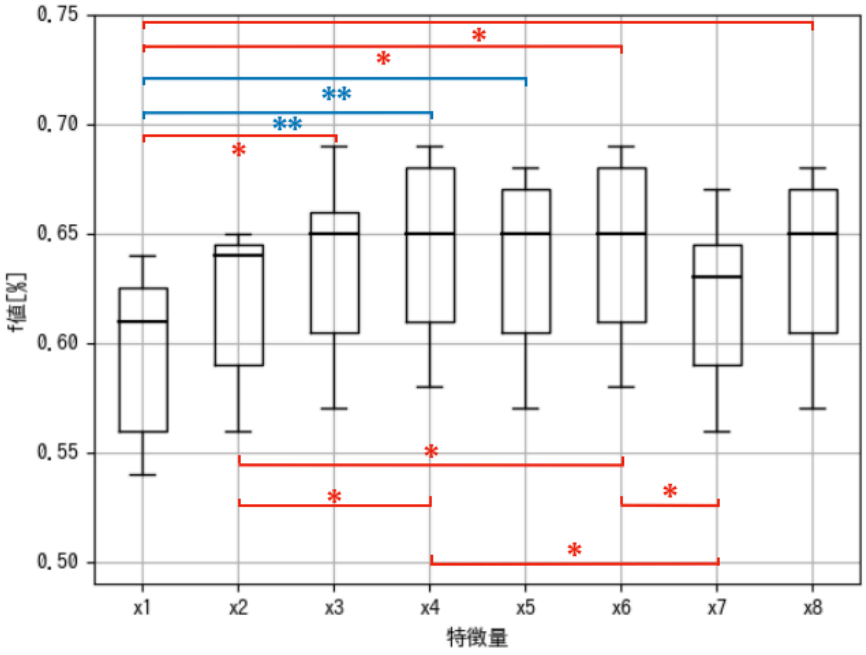
多重比較の結果、位相的特性から得られる情報に加えトピックモデルを考慮した特徴ベクトル $\boldsymbol{x}_4$ と $\boldsymbol{x}_1$ との間に有意差が認められた ($p < 0.01$)。さらに、$\boldsymbol{x}_4$ は、$\boldsymbol{x}_2$、$\boldsymbol{x}_7$ との間にも有意差が認められた ($p < 0.05$)。 いずれも、トピックモデルを考慮していない特徴ベクトルとの間に有意差が確認できた。

その他に $\boldsymbol{x}_1$ と $\boldsymbol{x}_3$、$\boldsymbol{x}_1$ と $\boldsymbol{x}_6$、$\boldsymbol{x}_1$ と $\boldsymbol{x}_8$、$\boldsymbol{x}_2$ と $\boldsymbol{x}_6$、$\boldsymbol{x}_6$ と $\boldsymbol{x}_7$ との間に有意差が認められた ($p < 0.05$)。 また $\boldsymbol{x}_1$ と $\boldsymbol{x}_5$ との間に有意差が認められた ($p < 0.01$)。

4.7 特徴量の重要度に関する評価

各特徴量を比較するために、ランダムフォレストを用いて特徴量の重要度を算出した。 その結果が図 4.4 である。 各特徴量の重要度は 0 から

図 4.3: 特徴ベクトルと f 値の箱ひげ図

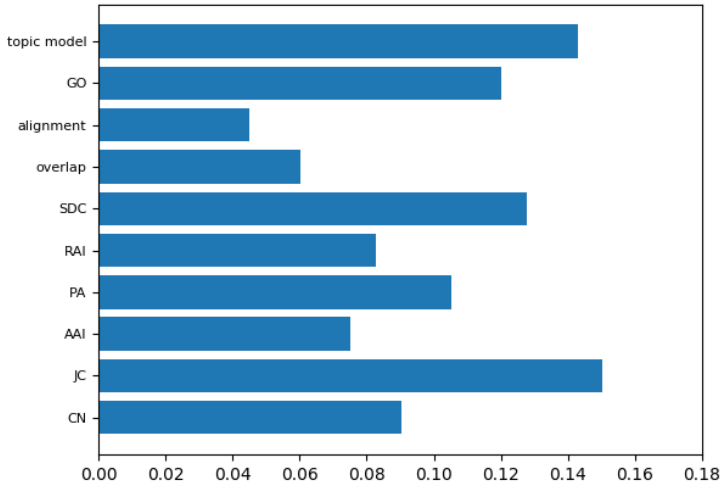


*：有意 ($p < 0.05$)、 **：有意 ($p < 0.01$)

1 の間の値をとり、総和が 1 になるように算出する。 重要度が 1 に近いほど重要度の高い特徴量であることを表す。

本実験の結果、*Jaccard's coefficient* が、最大の重要度となった (重要度 = 0.150)。 続いて、提案手法であるトピックモデルが重要であるという結果が得られた (重要度 = 0.143)。 その次に、*Sorensen Dice coefficient*、遺伝子オントロジーという順で重要であると判断できる (それぞれ、重要度 = 0.121、重要度 = 0.120)。 一方で、アミノ酸配列のアライメント情報は、2 番目に低い重要度となった (重要度 = 0.060)。 これらの結果から、今回の実験において、トピックモデルを考慮した特徴量は、分類をする上で重要な情報源となっていると考える。

図 4.4: 特徴量の重要度



CN : *common neighbors*、*JC* : *JAccard's coefficient*、*AAI* : *Adamic / Adar index*、*PA* : *Prefential attachment*、*RAI* : *Resource allocation*、*SDC* : *Sorensen dice coefficient*、*overlap* : *Overlap coefficient*、 *alignment* : アミノ酸配列のアライメント、*GO* : 遺伝子オントロジー、*topic model* : トピックモデル

4.8 議論

多重比較検定の結果、$\boldsymbol{x}_1$ と $\boldsymbol{x}_4$ との間に有意差が認められたことから、位相的特性から得られる情報に、提案手法であるトピックモデルを付加することで、タンパク質間の表現力を高め、精度向上に貢献できると考える。 さらに、既存手法と提案手法を組み合わせて構成される、$\boldsymbol{x}_5$ や $\boldsymbol{x}_6$、$\boldsymbol{x}_8$ に関しても、$\boldsymbol{x}_1$ と有意差が認められたことから、表現力を高め、精度向上に貢献できると考えることができる。

また、今回の実験では、遺伝子オントロジーを考慮した特徴ベクトルである $\boldsymbol{x}_3$ も $\boldsymbol{x}_1$ と有意差が認められた。 したがって、今回の実験では、遺伝子オントロジーも位相的特性から得られる情報に加えることで、表現力を高めることができると考える。 しかし、アミノ酸配列を考慮した特徴量である x_2 と x_1 との間に有意差は認められなかった。 さらに、x_4 と x_2 との間に有意差が認められたが、x_3 と x_2 との間には有意差が認められなかった。 このことから、遺伝子オントロジーを考慮した特徴量と

トピックモデルを考慮した特徴量の両方で表現力を高めているが、その貢献度はトピックモデルを考慮した特徴量の方が大きいと考える。

その他に、アミノ酸配列を遺伝子オントロジーに追加した特徴ベクトルである $\boldsymbol{x}_7$ は $\boldsymbol{x}_1$ との間に有意差が認められなかったのに対して、アミノ酸配列をトピックモデルに追加した特徴ベクトルである $\boldsymbol{x}_5$ は $\boldsymbol{x}_1$ と有意差があることから、トピックモデルは単体だけでなく、他の情報と共に利用する際にも表現力を高めることができると考える。

さらに、特徴量の重要度の評価を行なった結果から、トピックモデルを考慮した特徴量が、未知のタンパク質間相互作用において重要な情報源であると考えることができる。 このことからも、トピックモデルを考慮した特徴量が、タンパク質間の表現力を高める、精度向上に貢献できると考える。

4.9 今後の課題

本研究では、*STRING* と呼ばれるデータセットを用いて実験を行ない、本手法の有用性を確認した。 今後は、複数のデータセットにおいて本手法の有用性を評価し、汎用性を確認する必要がある。 さらに、今回用いた遺伝子オントロジーやアミノ酸配列の情報の他にも、遺伝子発現ベースの手法、比較ゲノムベースの手法、コドンベースの手法、さらに、本手法とは別の、自然言語処理やテキストマイニングの技術を用いて抽出される意味的情報など、様々な既存手法との比較や、組み合わせによる精度評価を行う必要がある。

第5章　関連研究

5.1　決定木

決定木 (*Decision Tree*) は、データの分類を繰り返し行い、分岐過程を階層化して木で表現される。現在、広く使われている *CART* (*Classification and Regression Tree*) 法は *Breiman* らによって提案された [27]。 *CART* 法は、分類だけでなく、回帰に対しても適用することができるアルゴリズムである。

決定木は、繰り返しデータを分割し、徐々にデータを分類していくが、データを分割するためには基準が必要である。 *CART* 法ではいくつかの基準が提案されているが、広く使われているのはジニ係数である。 ジニ係数は、経済学の領域で所得格差を表すのに用いられている指標であり、0 から 1 の間の値を取る。 値が 0 に近いほど平等であることを表す。 各分岐で、偏りが少ないほど、分岐が少ない決定木を構築することができるので、ジニ係数がもっとも低下するように分割していく。 その他にもエントロピーに基づく分類法がある。

5.2　ランダムフォレスト

ランダムフォレスト (*Random Forest*) は、*Leo* らによって提案されたアンサンブル学習 (*ensemble learning*) の一種である [28]。

ランダムフォレストは、複数の決定木を組み合わせて森 (*forest*) を構築することで識別、分類を行う。 個々に学習した複数の学習モデル (決定木) を組み合わせる事で汎化能力を向上させ、一つの学習モデルを作成する。 これはアンサンブル学習と呼ばれ、一つ一つの決定木は、アンサンブル学習における弱学習モデル (*weak classifier*) に相当する。

まず学習データを用いて、複数の決定木を構築する。 各決定木は、分岐ノード (*Split Node*) と葉ノード (*Leaf Node*) によって構成され、分岐ノードではそのノードの子のうちどれに進むべきかを決定する分岐関数

(*Split Function*) が与えられ、葉には最終的な出力結果が与えられる。また葉ノードは、そのノードに相当する学習データが持つクラス分布を与える。 識別は、未知データを各決定木に入力し、辿り着いた葉ノードに与えられているクラス分布を出力する。

5.3 k 最近傍法

k 最近傍法 (*k-Nearest Neighbor, k-NN*) は、回帰や分類を行う際に、類似するデータを k 個集めて多数決的に目的とする値を決定する手法である。 回帰の際には、類似する k 個のデータの平均値や中央値などの値を予測結果とする。 分類のときは、類似する k 個のデータが属するクラスのうち、もっとも数が多いクラスにデータを分類する。 k はパラメータで、事前に設定する必要がある。

5.4 ナイーブベイズ分類器

ナイーブベイズ分類器 (*naive bayes classifier*) は確率に基づいた分類器であり、事例 X に対して $P(c|X)$ が最大となるクラス $c \in C$ を出力する。このとき、ベイズの定理と呼ばれる次の性質を用いる。

$$P(c|X) = \frac{P(c)P(X|c)}{P(X)} \tag{5.1}$$

式 (5.1) の右辺を最大にする c が、出力となるクラスであるが、分母の $P(X)$ は c に依存しないので、最大にする c_{max} は次のように求める。

$$c_{max} = \arg\max_c \frac{P(c)P(X|c)}{P(X)} = \arg\max_c P(c)P(X|c) \tag{5.2}$$

ナイーブベイズ分類器は確率分布 $P(X|c)$ にシンプルな分布を仮定する。事例 X の正規確率を、正規分布を仮定して表すと、

$$P(X|c) = \prod_i P(x_i|c) = \prod_i \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp(-\frac{1}{2\sigma_c^2}(x_i - \mu_{c_i})^2) \tag{5.3}$$

となる。 したがって、ナイーブベイズ分類器の正規分布モデルは、

$$P(c)P(X|c) = P(c)\prod_i \frac{1}{\sqrt{2\pi\sigma_c^2}} \exp(-\frac{1}{2\sigma_c^2}(x_i - \mu_{c_i})^2) \tag{5.4}$$

を最大にする c を出力する。

5.5 ロジスティック回帰

あるデータ $\boldsymbol{x}$ が与えられたときのクラス C_1 の事後確率を、

$$p(C_1|\boldsymbol{x}) = \sigma(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x})) \tag{5.5}$$

と表すことができる。 ただし、$\sigma(\cdot)$ は式 (5.6) で定義されるロジスティックシグモイド関数、$\boldsymbol{w}$ は推定したいパラメータ、$\boldsymbol{\phi}(\cdot)$ は基底関数 $\phi_i(\cdot)$ を要素にもつ基底関数ベクトルである。 このパラメータ $\boldsymbol{w}$ の推定方法は様々だが、ここでは最尤法を用いて説明する。

$$\sigma(x) = \frac{1}{1+\exp(-x)} \tag{5.6}$$

最尤法なので、ロジスティックシグモイド関数を微分する。微分をすると、

$$\frac{d\sigma}{dx} = (1-\sigma(x))\sigma(x) \tag{5.7}$$

となる。データの集合　$\{\boldsymbol{x_n}, y_n\}$ が与えられたときの尤度は、

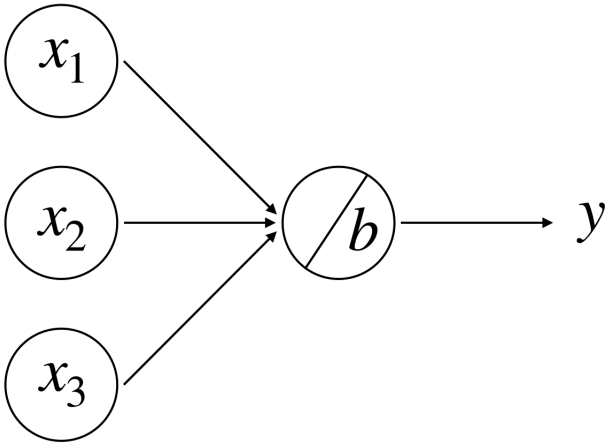
$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{w}) &= \prod_{n=1}^{N} p(C_1|\boldsymbol{\phi}(\boldsymbol{x_n}))^{y_n} p(C_2|\boldsymbol{\phi}(\boldsymbol{x_n}))^{1-y_n} \\ &= \prod_{n=1}^{N} p(C_1|\boldsymbol{\phi}(\boldsymbol{x_n}))^{y_n} \{1-p(C_1|\boldsymbol{\phi}(\boldsymbol{x_n}))\}^{1-y_n} \\ &= \prod_{n=1}^{N} \sigma(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})))^{y_n} \{1-\sigma(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n}))\}^{1-y_n} \end{aligned} \tag{5.8}$$

と表せる。 ただし、$y_n \in \{0,1\}$ である。 この負の対数尤度を、誤差関数として定義する。

$$\begin{aligned} E(\boldsymbol{w}) &= -\log p(\boldsymbol{y}|\boldsymbol{w}) \\ &= -\sum_{n=1}^{N} \{y_n \log \sigma(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})) \\ &+ (1-y_n)\log(1-\sigma(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x_n})))\} \end{aligned} \tag{5.9}$$

この誤差関数 $E(\boldsymbol{w})$ をパラメータ $\boldsymbol{w}$ で微分し、勾配が 0 に近くなるようにパラメータを調整していけば良い。 これには、ニュートン法や、勾配降下法などの手法を用いると良い。

図 5.1: ニューロンのモデル



5.6 ニューラルネットワーク

ニューラルネットワークは、人間におけるニューロン (神経細胞) の情報伝達の仕組みを模倣したものとして提案された。脳は、膨大な量のニューロンが集まって構成されていて、ニューロン同士が結合して脳の知的活動を実現している。

ニューラルネットワークでは、ニューロンを数理的にモデル化を行う。ニューロンは入力を受け取り、式 (5.10) を用いて値を計算し、その計算結果が設定した閾値を超えたとき、発火したものとみなし、他のニューロンに情報を伝達する。 また、ニューロンを視覚的に表現したものが、図 5.1 である。

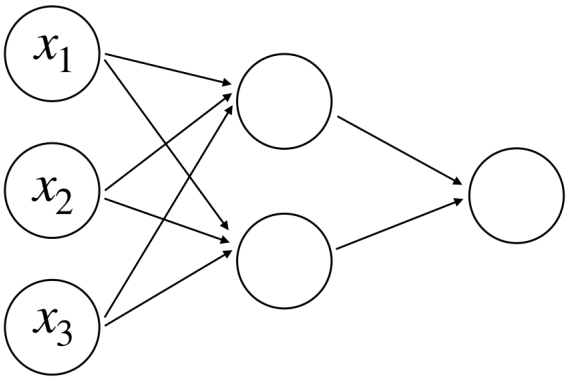
$$y_n = b + \sum_{d=1}^{D} w_{nd} x_d \tag{5.10}$$

ただし、n は全部で N 個のうちの n 番目のデータを表し、x_d は入力の d 次元目の要素を表す。 w_{nd} は重みであり、b はバイアスと呼ばれる。 y_n が閾値 θ を超えたときに 1 を、下回ったときに 0 を返す。

ニューラルネットワークは、ある層のニューロンが、次の層のニューロンに連結するように構成されたものである。 2 層のニューラルネットワークを視覚的に表現したものが、図 5.2 である。

学習時には、各ニューロン間の重みを出力層側から入力層側に向かって逆方向に調整していく。 この手法をバックプロパゲーション (*Backpropagation*) と呼ぶ。

図 5.2: 2 層のニューラルネットワーク



さらに、層を重ねることで、様々な目的に適応するモデルを構築することができる。 現在では深層学習 (*deep learning*) の研究が盛んであり、画像認識や自動運転、自然言語処理など様々な分野に応用されている。

5.7 遺伝子オントロジーに基づく類似尺度

今日までに、タンパク質の機能的類似性を定量化するために、遺伝子オントロジーに基づいた類似尺度が提案されている。 *Jain* らは、遺伝子オントロジーの階層構造に着目した類似尺度を提案している [29]。 a に付与された $GOterm_a$ と b に付与された $GOterm_b$ の共通の親に相当する $GOterm_p$ を見つけ、その $GOterm_p$ が、階層構造の深い位置で定義されているほど、類似度が高いという発想である。 図 5.3、図 5.4 では *GO term* の階層表現の例を表し、赤いノードが、着目している遺伝子に付与された *GO term* で、緑のノードが、それらの共通する親のノードである。図 5.3 の場合、緑のノードの深さが 2 であることに対し、図 5.4 の場合深さが 1 である。 したがって、図 5.4 における GO term 対に比べ、図 5.3 における *GO term* 対の方が類似度が高いことを表している。 各 *GO term* 対の類似度を算出した後に、その最大値を採用する手法、平均を採用する手法、共通する親の数を採用する手法などが提案されている。 その他にも *GO term* 対の類似性は、同一の遺伝子にアノテーションされている頻度によっても定義することができる [30]。

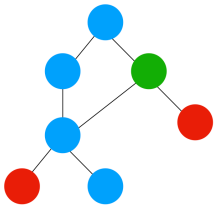


図 5.3: 例 1

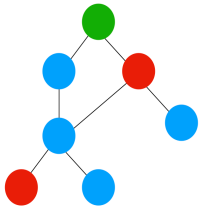


図 5.4: 例 2

5.8 文書間の類似度計算

文書間の類似度を計算するクラスタリングや分類のアルゴリズムは、ランキングアルゴリズム [31]、テキスト要約 [32]、機械翻訳 [33]、検索エンジン [34]、情報推薦 [35] といった技術の基礎的な手法として用いられる。

本節ではまず、第 5.8.1 項で、大量の文書から単語間の潜在的な類似度を測定するコーパスベースの手法を説明し、第 5.8.2 項で、*Web* 上のテキストが、他のページへのリンクを使用している点に着目した手法について述べ、最後に第 5.8.3 項で、*Web* 上の集合知である *WordNet* を利用する手法について述べる。

5.8.1 コーパスベース

コーパスベース類似度とは、文書に現れる各単語の類似性を大規模なコーパスから取得する手法である [36][37]。 *Mihalcea* らは単語の類似度と単語の特異性を用いて文書間の類似性尺度を定義している [38]。 単語の特異性は、逆文書頻度 (*Inverse Document Frequency, IDF*) によって決定される。 また、コーパスベースの手法として広く知られている手法の一つとして、潜在的意味解析 (*Latent Semantic Analysis, LSA*) がある。 *LSA* は意味が似ている単語は、類似している文書で出現する事を仮定している。 この仮定の下、まず、文書-単語行列を生成する。 この行列に対して、行列のランクを低減させる事によって、本来の単語間の類似性を保ちながら計算コストを低減できるように、行列のサイズを減らす特異値分解 (*Singuar Value Decomposition, SVD*) と、コサイン類似度を用いて文書間の類似度を求める。 すなわち、各行が各文書の特徴ベクトルを表すので、*SVD* によって次元が低減した特徴ベクトルを 2 つ取り出し、それらのコサイン類似度を求める事によって、効率良く文書の類似度を

測る事ができる [39][40][41]。 また *Kashyap* らは、*LSA* と *WordNet* を用いて生成された意味的単語類似度モデルを用いた *Semsim*[42] を提案している。 これはコーパスベースと *WordNet* を組み合わせた手法である。

5.8.2 リンクベース

Web テキストは、リンク情報を持っている場合が多いので、リンク情報を利用した類似度測定の手法が提案されている。 リンク情報を用いた類似度測定は、1 つのノードが 1 つの文書に対応し、ノード間の辺が文書間の類似度を表すグラフを用いて、文書間の類似度を測定するグラフデータマイニングが基礎となっている。 このようなグラフデータマイニングの基本的な手法として *SimRank*[43] が提案されている。 *SimRank* では、2 つのオブジェクトが類似しているオブジェクトと隣接していれば、その 2 つのオブジェクトは類似していると仮定している。 初期値は、同一オブジェクト間の類似度を 1、異なるオブジェクト間の類似度を 0 に設定する。 ノード間の類似度は再帰的に定義されていて、同一ノード間の類似度は 1、異なるノード間の類似度はそれぞれのノードが隣接する全てのノード間の類似度の平均で定義されている。 類似度計算は、収束するまで繰り返し計算される。 *Yoon* らは、あるオブジェクトから他のオブジェクトへの到達する確率を要素とするベクトルで各オブジェクトを表現し、2 つのオブジェクト間の類似度をそれぞれのベクトルのコサイン類似度によって計算する手法を提案した [44]。このようなリンクベースの手法は、ある閾値以上の類似度となるオブジェクトに対応するノード間にだけ、辺が存在するという仮定の下で定義されたグラフ構造であれば、どの分野にも適応可能である。

5.8.3 WordNet を用いた文書分類

WordNet とは、同義語の単語から 1 つのクラスタを作成し、論理的に類似しているクラスタ同士の関係を定義しているインターネット上の情報源である。 すなわち、字面は異なるが意味が同じ単語が同じグループにまとめられているので、単語間の意味的類似度を値とする特徴ベクトルを各文書に対して作成し、コサイン類似度を計算する文書分類の手法が提案されている [45][46][47]。 しかし、このような手法は計算量が膨大になる事が知られている [40]。 そこで *Madylova* らは精度を保ちながら計

算量を減らす事を目的とした手法を提案した [48]。 この手法は *is-a* 関係 [49] を満たすように分類された単語を下に単語の特徴ベクトルを構築し、その特徴ベクトルで文書を表現する。 その後、文書間の類似度をコサイン類似度を用いて算出するという手法である。 ここで *is-a* 関係を下に分類された単語間の意味的類似度は、*Wu* らが提案した *Wu-Palmer* 類似度 [50] を用いて算出している。

第6章　まとめ

本稿では、位相的情報に加え、タンパク質間の表現力を高め、精度向上に貢献するトピックモデルを考慮した新規特徴量を提案した。

本研究では、類似するタンパク質は、そのタンパク質に関して書かれている学術誌の潜在的トピックも類似しているという仮定の下に、各タンパク質の潜在的トピックを推定し、それらの潜在的トピックに対する類似度を新規特徴量として定義した。 今回潜在的トピックを推定するために、*LDA* を用いた。 位相的特性から得られる情報に加え、既存の手法であるアミノ酸配列の情報や、遺伝子オントロジーベースの情報と提案手法の差の検定を行い、各特徴ベクトル間の有意差を確認することで、提案手法の有用性を示した。 また、特徴量の重要度に関しても評価し、本手法の精度向上に対する貢献度を示した。

今後の課題としては、本手法の汎用性を示すために、複数のタンパク質間相互作用のデータセットに対して本手法を適用し、本手法の有用性を確認する必要がある。 また、本研究で使用した既存手法の他に、遺伝子発現ベースの手法、比較ゲノムベースの手法、コドンベースの手法、さらに本手法とは別の、自然言語処理やテキストマイニングの技術を用いた手法などとの比較や、それらの情報の組み合わせによって構成される特徴ベクトルの精度評価を行う必要がある。

謝辞

本研究を行うにあたりまして、お忙しい中多大なるご指導や研究の指針となる助言をしていただきました、東京理科大学情報科学科の滝本宗宏教授に心から感謝の意を表します。 また議論を通して、様々な知識や示唆を頂いた滝本研究室の皆様に感謝の意を表します。

最後に研究の場を与えてくださった東京理科大学に感謝の意を込めて、ここに厚く御礼申し上げます。

付録

本研究で作成したプログラムのソースコードを巻末に添付する。

参考文献

[1] Hui Ge, Albertha JM Walhout, and Marc Vidal. Integrating 'omic' information: a bridge between genomics and systems biology. *TRENDS in Genetics*, 19(10):551–560, 2003.

[2] Florian Iragne, Macha Nikolski, Bertrand Mathieu, David Auber, and David Sherman. Proviz: protein interaction visualization and exploration. *Bioinformatics*, 21(2):272–274, 2005.

[3] Damian Szklarczyk, John H Morris, Helen Cook, Michael Kuhn, Stefan Wyder, Milan Simonovic, Alberto Santos, Nadezhda T Doncheva, Alexander Roth, Peer Bork, Lars J. Jensen, and Christian von Mering. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Research*, 45(D1):D362–D368, 2017.

[4] Gary D. Bader and Christopher WV Hogue. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4(1):2, Jan 2003.

[5] Chengwei Lei and Jianhua Ruan. A novel link prediction algorithm for reconstructing protein–protein interaction networks by topological similarity. *Bioinformatics*, 29(3):355–364, February 2013.

[6] Fei Tan, Yongxiang Xia, and Boyao Zhu. Link prediction in complex networks: A mutual information perspective. *PLOS ONE*, 9(9):1–8, 09 2014.

[7] Q. Xu, E. W. Xiang, and Q. Yang. Protein-protein interaction prediction via collective matrix factorization. In *2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 62–67, Dec 2010.

[8] Lise Getoor and Christopher P. Diehl. Link mining: A survey. *SIGKDD Explor. Newsl.*, 7(2):3–12, December 2005.

[9] Lucy Skrabanek, Harpreet K. Saini, Gary D. Bader, and Anton J. Enright. Computational prediction of protein–protein interactions. *Molecular Biotechnology*, 38(1):1–17, Jan 2008.

[10] Maricel G. Kann, Benjamin A. Shoemaker, Anna R. Panchenko, and Teresa M. Przytycka. Correlated evolution of interacting proteins: Looking behind the mirrortree. *Journal of Molecular Biology*, 385(1):91 – 98, 2009.

[11] Hamed Shateri Najafabadi and Reza Salavati. Sequence-based prediction of protein-protein interactions by means of codon usage. *Genome Biology*, 9(5):R87, May 2008.

[12] Domonkos Tikk, Philippe Thomas, Peter Palaga, Jörg Hakenberg, and Ulf Leser. A comprehensive benchmark of kernel methods to extract protein–protein interactions from literature. *PLOS Computational Biology*, 6(7):1–19, 07 2010.

[13] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

[14] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. In *Proc. of 20th Intl. Conf. on VLDB*, pages 487–499, 1994.

[15] Paul Jaccard. Etude de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, 37:547–579, 01 1901.

[16] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25(3):211 – 230, 2003.

[17] A.L Barabási, H Jeong, Z Néda, E Ravasz, A Schubert, and T Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3):590 – 614, 2002.

[18] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, Oct 2009.

[19] T. Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons. *Biol. Skr.*, 5:1–34, 1948.

[20] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A.-L. Barabási. Hierarchical organization of modularity in metabolic networks. *Science*, 297(5586):1551–1555, 2002.

[21] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[22] S Henikoff and J G Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

[23] Robert T. Hersh. Atlas of Protein Sequence and Structure, 1966. *Systematic Biology*, 16(3):262–263, 09 1967.

[24] M. Ashburner. Gene ontology: Tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.

[25] Andreas Schlicker, Francisco S. Domingues, Jörg Rahnenführer, and Thomas Lengauer. A new measure for functional similarity of gene products based on gene ontology. *BMC Bioinformatics*, 7(1):302, Jun 2006.

[26] Yue Huang, Mingxin Gan, and Rui Jiang. Ontology-based genes similarity calculation with tf-idf. In *Proceedings of the Third International Conference on Information Computing and Applications*, ICICA'12, pages 600–607, Berlin, Heidelberg, 2012. Springer-Verlag.

[27] Olshen RA Stone CJ Breiman L, Friedman JH. Classification and regression trees. crc press; 1984. *CRC Press*, 1984.

[28] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[29] Shobhit Jain and Gary D. Bader. An improved method for scoring protein-protein interactions using semantic similarity within the gene ontology. *BMC Bioinformatics*, 11(1):562, Nov 2010.

[30] Xiaomei Wu, Erli Pang, Kui Lin, and Zhen-Ming Pei. Improving the measurement of semantic similarity between gene ontology terms and gene products: Insights from an edge- and ic-based hybrid method. *PLOS ONE*, 8(5):1–11, 05 2013.

[31] K. Sparck Jones, S. Walker, and S. E. Robertson. A probabilistic model of information retrieval: Development and comparative experiments. volume 36, pages 779–808, Tarrytown, NY, USA, November 2000. Pergamon Press, Inc.

[32] Mine Berker and Tunga Güngör. Using genetic algorithms with lexical chains for automatic text summarization. In *ICAART 2012 - Proceedings of the 4th International Conference on Agents and Artificial Intelligence, Volume 1 - Artificial Intelligence, Vilamoura, Algarve, Portugal, 6-8 Febrary, 2012*, pages 595–600, 2012.

[33] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318, 2002.

[34] Mehran Sahami and Timothy D. Heilman. A web-based kernel function for measuring the similarity of short text snippets. In *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*, pages 377–386, 2006.

[35] Ridvan Saraçoglu, Kemal Tütüncü, and Novruz Allahverdi. A fuzzy clustering approach for finding similar documents using a novel similarity measure. volume 33, pages 600–605, 2007.

[36] Yunus Emre Esin, Özgür Alan, and Ferda Nur Alpaslan. Improvement on corpus-based word similarity using vector space models. In *The 24th International Symposium on Computer and Information Sciences, ISCIS 2009, 14-16 September 2009, North Cyprus*, pages 280–285, 2009.

[37] Aminul Islam and Diana Zaiu Inkpen. Semantic text similarity using corpus-based word similarity and string similarity. volume 2, 2008.

[38] Rada Mihalcea, Courtney Corley, and Carlo Strapparava. Corpus-based and knowledge-based measures of text semantic similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence - Volume 1*, AAAI'06, pages 775–780. AAAI Press, 2006.

[39] C. A. Kumar and S. Srinivas. Latent semantic indexing using eigenvalue analysis for efficient information retrieval. In *International Journal of Applied Mathematics and Computer Science*, pages 551–558.

[40] Marcin Kuta and Jacek Kitowski. Comparison of latent semantic analysis and probabilistic latent semantic analysis for documents clustering. volume 33, pages 652–666, 2014.

[41] Hongjiao Xu, Wen Zeng, Jie Gui, Peng Qu, Xiaohua Zhu, and Lijun Wang. Exploring similarity between academic paper and patent based on latent semantic analysis and vector space model. In *12th International Conference on Fuzzy Systems and Knowledge Discovery, FSKD 2015, Zhangjiajie, China, August 15-17, 2015*, pages 801–805, 2015.

[42] Abhay L. Kashyap, Lushan Han, Roberto Yus, Jennifer Sleeman, Taneeya W. Satyapanich, Sunil R Gandhi, and Tim Finin. Robust Semantic Text Similarity Using LSA, Machine Learning and Linguistic Resources. volume 50, pages 125–161. Springer, March 2016.

[43] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 538–543, 2002.

[44] Seok-Ho Yoon, Ji-Soo Kim, Jiwoon Ha, Sang-Wook Kim, Minsoo Ryu, and Ho Jin Choi. Reachability vectors: features for link-based similarity measures. In *Symposium on Applied Computing, SAC 2014, Gyeongju, Republic of Korea - March 24 - 28, 2014*, pages 594–597, 2014.

[45] Mostafa Ghazizadeh Ahsaee, Mahmoud Naghibzadeh, and S. Ehsan Yasrebi Naeini. Semantic similarity assessment of words using weighted wordnet. volume 5, pages 479–490, 2014.

[46] Hongzhe Liu and Pengfei Wang. Assessing sentence similarity using wordnet based word similarity. volume 8, pages 1451–1458, 2013.

[47] Shen Wan and Rafal A. Angryk. Measuring semantic similarity using wordnet-based context vectors. In *Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Montréal, Canada, 7-10 October 2007*, pages 908–913, 2007.

[48] Ainura Madylova and Sule Gündüz Ögüdücü. A taxonomy based semantic similarity of documents using the cosine measure. In *The 24th International Symposium on Computer and Information Sciences, ISCIS 2009, 14-16 September 2009, North Cyprus*, pages 129–134, 2009.

[49] Zhibiao Wu and Martha Stone Palmer. Verb semantics and lexical selection. In *32nd Annual Meeting of the Association for Computational Linguistics, 27-30 June 1994, New Mexico State University, Las Cruces, New Mexico, USA, Proceedings.*, pages 133–138, 1994.

[50] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32Nd Annual Meeting on Association for Computational Linguistics*, ACL '94, pages 133–138, Stroudsburg, PA, USA, 1994. Association for Computational Linguistics.