Vol.2012-BIO-29 No.15
2012/6/28

複数の分類器に基づく半教師あり学習を用いた文献からの蛋白質間相互作用抽出

小薮 駿[1,a)]　大川 剛直[1]

概要：文献からのタンパク質相互作用情報抽出において，十分な量の訓練データが得られない場合，仮ラベル推定に基づく半教師あり学習が有効である．このようなタイプの半教師あり学習では仮ラベルを与える際に，誤ってラベルを付与することが精度低下の原因となるため，いかに正確に仮ラベルを付与するかが，極めて重要である．そこで本研究では、複数の分類器を用い，その共通コンセンサスを得る際に，分類器の類似度や学習手法の信頼度を導入することで正確な仮ラベル決定が可能となる手法を提案する。相互作用情報抽出実験の結果として，データセットが比較的大きな場合に，提案手法を用いることで，より精度の高い抽出が達成された．また従来手法との比較において，F 値と再現率では同等，もしくは少し劣る結果となったが，適合率の観点では提案手法が優位な結果を示すことが確認された．

キーワード：情報抽出，蛋白質間相互作用，半教師あり学習

Extracting protein-protein interaction from litratures based on semi-supervised learning using multiple classifiers

Shun Koyabu[1,a)]　Takenao Ohkawa[1]

Abstract: Semi-supervised learning based on tentative label prediction is a useful technique for automatic extraction of protein-protein interaction from litratures if enough training instances cannot be prepared. In such a framework of semi-supervised learning, how we predict the correct labels is very important for accurate extraction. In this paper, we propose a method of predicting tentative labels based on multiple classifiers introducing two types of measures for evaluating each classifier, similarity among the classifiers and reliability of the classifiers. As a result of experiment, the proposed method shows higher precision values for relatively large dataset, in comparison with conventiional methods.

Keywords: information extraction, protein-protein interaction, semi-supervised learning

1. はじめに

蛋白質相互作用に関する情報は，様々な医学生物学文献に記述されているが，このような文献は膨大に存在し，日々その数は増え続けている．そのため手作業で網羅的に相互作用する蛋白質のペアを特定することは時間的にも，労力的にも困難な状況に陥っている [1]．そこで，文献から計算機処理により蛋白質間相互作用情報（PPI）を自動抽出する手法に関する研究が進められている [2]．自動抽出のアプローチとして，PPI に関しての情報が判明しているものを訓練データとして使用した教師あり学習が広く用いられている．しかしながら，訓練データが常に豊富に存在するとは限らず，また新たな訓練データ作成には多大なコストがかかる可能性もある．このような状況においてはラベル（PPI の有無）が特定されていないデータも訓練に使用する半教師あり学習 [3] が有効である．半教師あり学習においてラベル未知データを利用する１つの方法として，ラベル未知データに擬似ラベルを与える方法が挙げられる．

[1] 神戸大学大学院システム情報学研究科
1-1 Rokkodai, Nada, Kobe, 657–8501, Japan
[a)] s.koyabu@cs25.scitec.kobe-u.ac.jp

こういった半教師あり学習においては信頼出来る擬似ラベルを付与することが重要となる．そこで複数の学習手法を利用した Multiview Learning により信頼性の高い擬似ラベルを与える試みがある [4]．一般に，Multiview Learning の枠組において擬似ラベルを付与する際には，各学習手法によるラベルの予測結果をもとに，その多数決をもって共通のコンセンサスとする方法が多く用いられる．しかしながら，このような方法の場合，非常に信頼性が高い学習手法が予測した結果であっても，それが少数意見であるならば採用されないという問題がある．そこで，共通コンセンサスを導く際に，分類器の類似度，学習手法の信頼度という尺度を導入し，それに基づいて半教師あり学習を行う手法を提案する．

2. 機械学習による蛋白質間相互作用情報抽出

2.1 蛋白質間相互作用情報とその機械学習による自動抽出

蛋白質の機能解明の上で蛋白質の結合，相互作用情報は必要不可欠な情報である．この中でも特に蛋白質と蛋白質の間での相互作用を，蛋白質間相互作用（PPI）と呼ぶ．蛋白質について述べられた文献には，疎水性であるかどうかといった蛋白質自身の特性に関する情報，蛋白質が発現する機能情報，並びに PPI に関する情報といった多様な情報が記述されている．このような文献の各文に存在する蛋白質の２つ１組の組み合わせ（蛋白質ペア）に対して，相互作用が認められているような蛋白質ペアを相互作用ペアと呼び，抽出すべき PPI と考える．以下の文 (1)，(2) を対象に，相互作用ペアの例を示す．

(1) GerE binds to a site on one of these promoters, cotX, that overlaps its -35 region.

(2) IL-6 promotes coprecipitation of p85 with gp130, the signal-transducing component of the IL-6 receptor.

これらは文献中に見られた複数の蛋白質名が登場する文である．文 (1) に関しては，蛋白質 “ GerE ” と “ cotX ” がある特定の部位で結合し，実際に相互作用を行うペアであることが示されているため，相互作用ペアとみなすことができる．また文 (2) にように “ IL-6 ”, “ gp130 ”, “ IL-6 receptor ” と３つ以上の蛋白質が文中に存在することもあり，この文の場合，３通りの蛋白質ペアを想定することができる．実際には, “ IL-6 ” と “ gp130 ” のペア間には相互作用が存在するが，他の２例, “ IL-6 ” と “ IL-6 receptor ”, ならびに “ gp130 ” と “ IL-6 receptor ” の間には相互作用が成立しないため，これらは相互作用ペアとはならない．

PPI を有するペアを正例（positive），そうでないペアを負例（negative）として扱い，２クラスの分類問題として考える．そして，既に正であるか負であるかが判別している事例（蛋白質ペア）をもとに，その蛋白質ペアの文中での記述上の特徴を用いて，正と負を区別する基準を自動的に見つけることが，機械学習による PPI 抽出の基本的な考え方となる．

2.2 蛋白質ペアの特徴

PPI に関して文中から得られる特徴として，相互作用を直接的に表現する記述であったり，それを暗示するような単語の存在であったり，また逆に相互作用がないことの記述であるなど，さまざまな記述上の特徴が挙げられる．そこで，分類の手掛かりとするため，これらの記述や単語等を各蛋白質ペアに対して特徴として付与する．以下にその具体例を示す．これらの特徴は既存の PPI 抽出の研究の多くで使用されており [5][6]，本研究でもそれらに準じて特徴を設定する．

2.2.1 文から得られる特徴を用いた属性

- **蛋白質ペアに関連する単語** (keyword)

２つの蛋白質の関係を表している単語を特徴として使用する．蛋白質ペアが記述されている文中から一般に相互作用を記述するときに頻繁に使用される *“interact”*，*“bind”*，*“active”*，*“depend”* などの単語 642 種類を指定する．keyword として抽出された単語を原型に戻し，ステミングを行った 180 種類を特徴として使用する．keyword となる単語の候補が複数ある場合は，対象蛋白質ペアの間に存在するものを優先して，蛋白質との単語間距離（下記の項を参照）が近いものを採用する．

単一文中に複数の蛋白質ペアが存在する場合はそのそれぞれに対し keyword が設定される．

- **蛋白質ペアと keyword の単語間距離**

対象蛋白質と keyword の間の距離によって語の関係性の強弱を評価する．単語間距離には２種類あり，蛋白質ペアのうち文に先に現れる蛋白質を A，後に現れる蛋白質を B とすると，A・B・keyword の３つ組において，先に現れる２つの間の単語間距離と，後に現れる単語間距離の２種類が定義される．

- **蛋白質ペアと keyword の順序**

蛋白質ペアと keyword の語順を特徴として用いる．蛋白質ペアを A・B，keyword を K とすると，文章中に現れる順序によって ABK, AKB, KAB の３種類に決定する．

- **蛋白質ペアと keyword 間のコンマの有無**

コンマというのは，しばしば文章の切れ目や，物事の列挙に際して使われるため，その前後で話題が変わっていることが多い．蛋白質ペア，keyword を文頭からの出現順に A, B, C とすると，AB 間のコンマの有無と BC 間のコンマの有無の組み合わせとして 4 種類の値を付与する．

- **否定語の有無**

文章に否定語が入っていると keyword の意味が否定され逆の意味になる．そのため蛋白質ペアの間，もしくは keyword と蛋白質の間に *“not”*，*“unable”*，*“incapable”* などの否定語が入っているかどうか特徴として使用する．

- **文の接続関係を表す単語の有無**

接続詞や, 接続副詞など, 文の前後の連接関係を表す語は, コンマと同様にその前後で話題が変わることがしばしばある．そのため蛋白質ペアの間, もしくは keyword と蛋白質の間にこれらが存在するかどうかを特徴として使用する．文の接続関係を表す単語として使用するのは, *"where"*, *"when"*, *"what"*, *"why"*, *"how"*, *"as"*, *"though"*, *"although"*, *"because"*, *"so"*, *"therefore"*, *"hence"*, *"since"*, *"wherein"*, *"whereas"*, *"whereby"* の 16 種類である．

- **which の有無**

蛋白質ペアの間，もしくは keyword と蛋白質の間に "which" が存在するかどうかを特徴として使用する．which は上記と同様に接続関係を表す単語であるが，上記の単語に比べ頻出でかつ使用される場合は比較的前後の関係を切る意味で使用されることが多いため，これらとは区別して取り扱う．

- **but の有無**

but も which と同様に接続関係を表す単語の中でも頻出で，その意味が前後関係の否定であり特殊である．蛋白質ペアの間，もしくは keyword と蛋白質の間に "but" が存在するかどうかを特徴として使用する．

- **仮定・条件を表す単語の有無**

仮定・条件を表す単語として *"if"*, *"whether"* を用いる．蛋白質ペアの間，もしくは keyword と蛋白質の間にこれらの単語が存在するかどうかを特徴として使用する．

- **keyword の前置詞**

keyword の後に前置詞がつくことによって意味が変化することがしばしばある，そこで keyword に続く前置詞そのものを特徴として用いる．但し，単語間距離で 3 以内に限定し，複数存在する場合は単語間距離が近いものを採用する．

- keyword **の出現回数**

特徴として使用する keyword は蛋白質ペア 1 つに対して 1 単語であるが，keyword になりえる単語が文中に複数存在する場合は，文自体が比較的強く相互作用に関連していることが想定される．このことから文中に複数の keyword が存在するかどうかを特徴として使用する．

2.2.2 構文解析情報から得られる特徴を用いた属性

蛋白質ペアについての記述がある文を構文解析器 (Stanford parser[7]) に通し，構文木を作成する．その構文木から得られる特徴を使用する．

- **蛋白質ペア・keyword の構文木における高さ**

蛋白質ペアならびに keyword の構文木での高さ（深さ）をそれぞれ特徴として使用する．これにより対象となる蛋白質ペアと keyword の単語間距離とは異なる近さや階層構造を表現できる．蛋白質ペアの 1 つ目の蛋白質の高さ，2 つ目の蛋白質の高さ，また蛋白質ペアに設定された keyword の高さを特徴として使用する．

- **蛋白質ペア・keyword の構文木における品詞情報**

蛋白質ペアならびに keyword の構文木での PATH（通り道）の品詞情報を特徴として使用する．これにより対象となる蛋白質ペアと keyword の構文構造を表現でき，分類器に擬似的な文法構造を学習させることができる．蛋白質ペアの 1 つ目の蛋白質を表す葉についての，根からの PATH, 同様に 2 つ目の蛋白質の PATH，また蛋白質ペアに設定された keyword の PATH を特徴として使用する．

3. 複数の分類器に基づく半教師あり学習

3.1 半教師あり学習の導入

精度の良い学習を行うためには，様々な情報を持つ多くのクラスラベル付きデータ（以下，既知データ）で分類器を十分に学習することが必要である．しかし既知データの作成には，専門家による仔細な調査や，時間的労力のため，豊富に用意することは難しい．一方でクラスラベルが未知のデータ（以下，未知データ）は，一般に多数存在し，獲得が容易である．既知データが十分ではなく，未知データが多数存在する場合に，半教師あり学習という手法が有効である．

半教師あり学習の 1 つのアプローチとして，訓練データ（既知データ）で分類器を訓練し，その訓練結果を用いて未知データの擬似的なクラスラベルの値（以下, 仮ラベル）を予測して与える方法がある．この方法では，仮ラベルを与えた未知データを訓練データに追加し，以前より大きな訓練データで訓練することを繰り返すことで性能向上を実現する．しかしながら，このようなタイプの単純な半教師あり学習では，分類器自身が分類した結果に基づいて仮ラベルを付与するため, 分類器自身が知っていること（分類器が判断したラベル）を再学習することになり, 本質的な性能の向上を見込むことは難しい．そのため複数個の分類器を作成し, 互いに学習結果を教えあう Co-training や MultiView Learning[4] と呼ばれる手法も半教師あり学習の拡張としてよく用いられる．本研究では，MultiviewLearning の考え方をもとに，複数の学習手法を用いて複数個の分類器を作成して未知データの仮ラベルを予測する際に，分類器間の類似性や分類器の信頼度をもとに，分類結果の採否をコントロールする仕組みを導入することで，仮ラベル付与を限定しより正確な仮ラベル付けが行えるような手法を提案する．

3.2 複数の分類器を用いた仮ラベルの推定

3.2.1 仮ラベル推定手法

半教師あり学習においては，仮ラベル付与の際にその事例が，正確な分類結果であること，すなわち，分類結果の信頼性が非常に重要である．なぜならば，仮ラベル付与に際し，誤ったラベルを多数付与して訓練データに追加すると，うまく学習が働かず，結果として精度低下を招く恐れがあるからである．そこで分類結果の信頼性が高い事例の

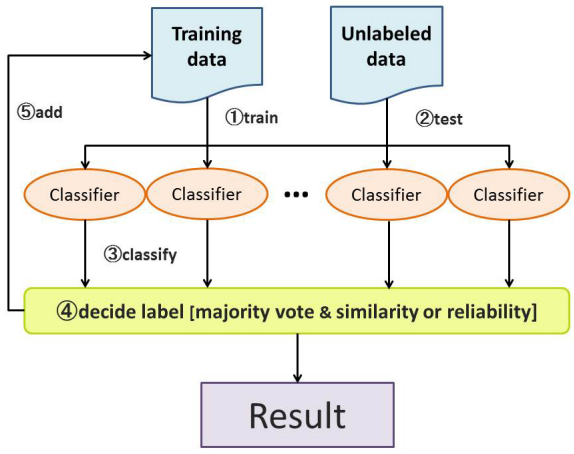

図 1 提案手法の概観
Fig. 1 General flow of proposed method

みを訓練データに追加する必要がある．仮ラベルを付与する事例の選定方法として，本研究では，多様な種類の学習手法を用意し，多数決を用いた方法を基本とする．このとき単に訓練データの特徴を分割することで複数の分類器を学習するのではなく，学習に用いる学習手法を変えることで多様な分類器を生成する．学習手法の異なる複数の分類器の間で共通に判断された事例は，多様な側面においてラベルの予測結果が一致したと考え，信頼性の高い事例であると判断できる．

対象とする事例に分類器がうまく適合するかどうかは，使用した学習手法に依存する．この適合の善し悪しのために単純な多数決では，誤りの少ない仮ラベル付けが可能とは限らない．なぜならば，比較的うまく適合している少数の分類器による予測結果が，多数のあまり適合していない分類器の存在によって，覆される可能性があるからである．そこで本研究では分類器間の類似度，学習方法の信頼度という学習結果の適合の善し悪しを評価するための尺度を導入し，うまく適合する分類器を考慮することで，訓練データに追加する事例を限定することにより，誤った仮ラベルの付与を抑止する．具体的には分類器間の類似度や学習方法の信頼度が最も高い値を示す分類器と多数決の判定が一致した場合のみ仮ラベル付与を行い訓練データに追加する．提案手法の全体の流れを Fig.1 に示す．

未知データから訓練データへ追加する事例を選択するアルゴリズムの擬似コードを Fig. 2 に示す．ここで，$\mathbf{S} = \{S_1, S_2, ..., S_M\}$ は未知データ集合，S_i は未知データの事例，M は未知データ数，C_i は分類器，N は分類器の総数，$\mathbf{C} = \{C_1, C_2, ..., C_N\}$ は分類器群をそれぞれ表す．また $f(C_i, \mathbf{C})$ は分類器群 $\mathbf{C}$ に対する分類器 C_i の類似度，$g(C_i)$ は分類器 C_i に使用されている学習方法の信頼度を与える関数であり，詳細は次節にて説明する．基本的には未

Procedure : add unlabeled data to training set
1 for($k = 1..k_{max}$)
2 Train($\mathbf{C}$) with training set.
3 $\mathbf{R}_k = \mathbf{C}$.classify($\mathbf{S}$).
4 for(i = 1..N)
5 $e_i = f(C_i, \mathbf{C})$ or $g(C_i)$.
6 for(i = 1..M)
7 $label_i = \mathbf{C}$.majority_decision(S_i).
8 $C = C_i$ such that $e_i == \max(e_1 .. e_N)$.
9 if($label_i == C$.classify(S_i))
10 add S_i with $label_i$ to training set.
11 if($\mathbf{R}_k == \mathbf{R}_{k-1}$) break;

図 2 未知ラベルデータの訓練データへの追加手順
Fig. 2 Procedure for adding unlabeled data to training set

知データ集合 $\mathbf{S}$ の全ての分類結果が前回の繰り返し時の結果と同一になった時点で収束したと判断し，処理を終了するが，繰り返し回数 k の上限値 k_{max} を設定することにより，一定回数で学習を打ち切るものとする．

3.2.2 分類器の類似度

分類器の類似度は，全ての未知データに対して分類器の判断が他の種類の分類器とどの程度一致しているかの指標である．すべての分類器が全ての未知データに対して全て同じ判断を下した場合が最大となるような値となる．この指標を使用する理由は，分類器が他の分類器と共通なコンセンサスをとれている時，この分類器の信頼性が高いと考えるためである．

Ω を全未知データ集合とし，A を未知データ（$\in \Omega$）とする．分類器群 $\mathbf{C}$ に対する分類器 C_i の類似度 $f(C_i, \mathbf{C})$ を以下のように定義する．

$$f(C_i, \mathbf{C}) = \sum_{A \in \Omega} (P(A, C_i))$$

但し，

$$P(A, C_i) = \frac{A \text{ に対して } C_j \text{ と同一のラベルを予測した分類器数}}{N}$$

である．

3.2.3 学習方法の信頼度

訓練データの中で k-folds Cross Validation（CV，交差検定法）により事前学習 (PreTraining) を行うことで，入力された訓練データにおける学習方法の信頼度を評価する．この指標を使用する理由は，与えられた訓練データと各分類器が使用している学習手法の信頼度が大きい程，精度の良い分類を行う分類器であると考えるためである．

学習方法の信頼度 $g(C_i)$ は分類器 C_i に使用されている学習方法を用いて以下のように定義される．ここで，$F(C, T, E)$ は，訓練データ T を用いて分類器 C を学習し，テストデータ E に対して求めた F 値を，W は事前学習に用いるすべての訓練データ集合，w_i は W を k 等分した各

集合を表す.

$$g(C_i) = 1/k \sum_j F(C_i, W - w_j, w_j)$$

4. 評価及び考察

4.1 評価実験

評価実験では既にクラスラベルが分かっている文書集合（コーパス）を利用し，2.2 で述べた蛋白質ペアに関する文の特徴を用いて評価用データを作成した．サイズが異なる 3 種類のコーパス，LLL，HPRD50，IEPA の 3 つ [8] を利用した．各コーパスの文数，蛋白質ペア，相互作用ペアの総数を Table 1 に示す.

用意する学習手法は Table 2 に示す 8 種類である．分類器にはデータマイニングツールボックス Weka に含まれているものを使用した [9].

評価実験 1：単一の学習方法（Baseline 1） 8 つの学習方法をそれぞれ単独で使用する手法.

評価実験 2：８つの学習手法（Baseline 2） 8 つの学習手法を組み合わせて結果をマージする手法.

評価実験 3：半教師あり学習手法（Baseline3） ８つの学習手法の単純な多数決により仮ラベルを決定する半教師あり学習手法.

評価実験 4：類似度を用いた提案手法（Proposed 1） 3.2.2 で示した分類器の類似度を用いた提案手法.

評価実験 5：信頼度を用いた提案手法（Proposed 2） 3.2.3 で示した分類器の信頼度を用いた提案手法.

半教師あり学習における繰り返し回数の上限（k_{max}）は 20 回とする．なお，評価実験 4, 5 においては，最終的な評価の際に 8 つの分類器の判断が二分した場合は学習方法の信頼度の最も高い値を持つ分類器の判断を採用する.

評価方法としては，Table 1 の各コーパスから作成した評価用データにおいて 10-folds CV で訓練データとテストデータにわけ，それぞれのテストデータに対する再現率，適合率，F 値の平均を評価に用いる．また，半教師あり学習における未知データは，10-folds CV におけるテストデータとする.

表 1 コーパス一覧
Table 1 Statistics on corpora

Corpus	LLL	HPRD50	IEPA
PPI pairs	164	163	335
All pairs	330	433	817

表 2 実験に使用した学習手法
Table 2 Learning methods for generating classifiers

Support Vector Machine(SVM)[10]
C4.5[11]
RotationForest(RotationF)[12]
KStar[13]
RandomForest(RandomF)[14]
CART[15]
Decorate(Deco)[16]
AdaBoost(AB)[17]

4.2 考察

評価実験 1 の各コーパスにおける実験結果を Table 3〜5 に示す．なお，太字は同じ評価項目の最大の値を表す．各結果において，学習方法の選択により，様々な精度を示すことが見て取れる．また与えられるデータセットに対する適合の良し悪しがあることがわかる．例えば CART アルゴリズムにおいて，LLL コーパス，IEPA コーパスでは他の学習方法と比べ，比較的精度が悪いが，HPRD50 コーパスにおいてはデータセットにうまく適合し，最も高い精度を示している．これにより学習手法毎に様々な傾向を持つ分類器が作成されていることがわかる.

次に，評価実験 2 の結果を Table 6 に示す．ここで Average F-Score of Beseline1 は評価実験 1 における F 値の値の平均値，Max F-score of baseline1 は評価実験 1 における最大の F 値である．８つの学習手法を組み合わせて結果をマージする手法での F 値は，LLL コーパスの時，評価実験 1 の最大 F 値を上回った．一方で，HPRD50 コーパス，IEPA コーパスにおいては最大 F 値を下回った．これは学習手法には前述のとおりうまく働くデータセットが存在するからである．しかし現実問題としてデータセットごとに

表 3 LLL: 実験結果, (Baseline 1)
Table 3 LLL: Experimental results(Baseline 1)

	SVM	C4.5	RandomF	RotationF	CART	KStar	Deco	AB
F-Score	0.762	0.714	0.770	**0.793**	0.712	0.764	0.764	0.731
Recall	0.780	0.738	0.805	0.805	0.732	0.780	**0.811**	0.744
Precision	0.744	0.691	0.737	**0.781**	0.694	0.749	0.723	0.718

表 4 HPRD50: 実験結果, (Baseline 1)
Table 4 HPRD50: Experimental results (Baseline 1)

	SVM	C4.5	RandomF	RotationF	CART	KStar	Deco	AB
F-Score	0.673	0.647	0.719	0.729	**0.734**	0.727	0.691	0.614
Recall	0.669	0.607	0.730	0.736	0.755	**0.791**	0.699	0.577
Precision	0.677	0.692	0.708	**0.723**	0.715	0.672	0.683	0.657

表 5 IEPA: 実験結果, (Baseline 1)
Table 5 IEPA: Experimental results (Baseline 1)

	SVM	C4.5	RandomF	RotationF	CART	KStar	Deco	AB
F-Score	0.633	0.583	0.647	**0.663**	0.620	0.634	0.647	0.625
Recall	0.615	0.481	0.633	0.636	0.624	**0.687**	0.636	0.636
Precision	0.652	**0.742**	0.663	0.692	0.617	0.588	0.659	0.614

表 6 実験結果, (Baseline 2)
Table 6 Experiment results, (Baseline 2)

Corpus	LLL	HPRD50	IEPA
F-Score	0.794	0.722	0.645
Recall	0.811	0.693	0.582
Precision	0.778	0.751	0.722
Average *F*-Score of Beseline 1	0.751	0.692	0.632
Max *F*-score of Baseline 1	0.793	0.734	0.663

表 7 実験結果, (Baseline 3)
Table 7 Experiment results, (Baseline 3)

Corpus	LLL	HPRD50	IEPA
F-Score	0.824	0.729	0.660
Recall	0.854	0.718	0.600
Precision	0.795	0.742	0.734

うまく適合する学習手法を発見することは難しく，必ずしもうまく働く学習手法を選択できるとは限らない．故に多くの学習手法の平均精度に比べ，比較的よい精度を示した複数の分類器を使用する学習手法は有用であるといえる．

次に評価実験 3 の実験結果について Table 7 に示す．半教師ありの学習手法を使用することで，使用していない評価実験 2 の結果よりも，*F* 値で，LLL コーパスにおいて 0.3 ポイント，HPRD50 コーパスにおいて 0.07 ポイント，IEPA コーパスにおいて 0.15 ポイント高い精度を示した．再現率，適合率においても同様に精度が上昇した．これは未知データとしての扱いであるテストデータの仮ラベル予測を行い擬似的に訓練データ数を増やしたため，より豊富な訓練データで分類器を学習でき，半教師あり学習を用いない場合と比べて効果的に学習がなされたためと考えられる．

さらに，提案手法である，評価実験 4，評価実験 5 についての実験結果を，それぞれ Table 8，9 に示す．HPRD50 コーパス，IEPA コーパスにおいて，それぞれ評価実験 3 に比べ，精度の向上が確かめられた．これは半教師あり学習で仮ラベルを決定する際に，うまく信頼性のある事例を判断し，間違った仮ラベルの付与件数を減らすことができたためであると考えられる．一方，LLL コーパスでは若干の精度低下がみられた．これは LLL コーパスは他のコーパスに比べ蛋白質ペア事例が少ないため事例のバリエーションの絶対数が少なく，半教師あり学習の目的である様々な事例を用いた訓練データでの学習という点においてうまく働かなかったためであると考えられる．よって今回の提案手法においてはデータセットが比較的大きい場合に有用であると考えられる．また類似度を用いた提案手法と，信頼度を用いた提案手法を比較すると，*F* 値と Recall については信頼度を用いる場合のほうが比較的有効に働いている．しかし Precision に関しては，類似度を用いると，LLL コーパスで同等，IEPA コーパスに関しては比較的有効に働いている．ゆえに適合率が必要となる場合，すなわち抽出判断の誤りを減らしたい場合に有用である．

また，コーパス毎の評価実験毎の比較を Table 10〜12 に示す．なお，評価実験 1（Baseline1）は各学習手法の平均値を表示している．

LLL コーパスにおいては，上述した通り，提案手法で若干の精度の低下がみられる．しかし Precision に関しては，ベースラインとなる評価実験に比べ上昇している．これは事例のバリエーションが少なく再現性を取ることはできなかったが，繰り返し学習することで適合率を上昇させ

表 8 実験結果, (Proposed 1)
Table 8 Experiment results, (Proposed 1)

Corpus	LLL	HPRD50	IEPA
F-Score	0.819	0.730	0.663
Recall	0.841	0.712	0.603
Precision	0.798	0.748	0.737

表 9 実験結果, (Proposed 2)
Table 9 Experiment results, (Proposed 2)

Corpus	LLL	HPRD50	IEPA
F-Score	0.819	0.738	0.669
Recall	0.841	0.724	0.615
Precision	0.798	0.751	0.733

表 10 LLL: 実験手法毎の比較
Table 10 LLL: Summary of experiments

Corpus	Beseline 1	Baseline2	Baseline3	Proposed 1	Proposed 2
F-Score	0.751	0.794	**0.824**	0.819	0.819
Recall	0.774	0.811	**0.854**	0.841	0.841
Precision	0.730	0.778	0.795	**0.798**	**0.798**

表 11 HPRD50: 実験手法毎の比較
Table 11 HPRD50: Summary of experiments

Corpus	Baseline1	Baseline2	Baseline3	Proposed 1	Proposed 2
F-Score	0.692	0.722	0.729	0.730	**0.738**
Recall	0.696	0.693	0.718	0.712	**0.724**
Precision	0.691	0.751	0.741	0.748	**0.752**

表 12 IEPA: 実験手法毎の比較
Table 12 IEPA: Summary of experiments

Corpus	Baseline1	Baseline2	Baseline3	Proposed 1	Proposed 2
F-Score	0.632	0.645	0.660	0.663	**0.669**
Recall	**0.619**	0.582	0.600	0.603	0.615
Precision	0.653	0.722	0.734	**0.737**	0.733

表 13　従来手法と提案手法の比較
Table 13　Comparison between method and proposed method

Bui's method	LLL	HPRD50	IEPA
F-Score	**0.841**	**0.738**	**0.747**
Recall	**0.841**	**0.779**	**0.839**
Precision	**0.841**	0.702	0.674
Proposed method	LLL	HPRD50	IEPA
F-Score	0.819	**0.738**	0.669
Recall	**0.841**	0.724	0.615
Precision	0.798	**0.752**	**0.733**

ることができたと考えられる．HPRD50 コーパスにおいては，提案手法 2（Proposed 2）を用いた場合，*F* 値，Recall，Precision の全てにおいて上昇した．IEPA コーパスにおいては，両提案手法で *F* 値において提案手法の有用性が確かめられた．一方で Recall は単純な多数決で半教師あり学習を用いた手法（Baseline3）に比べると上昇が示されているが，各学習方法を単独で使用した場合の平均値に比べると低下している．これは IEPA コーパスに対する各学習手法の精度が，LLL コーパスで 0.75 程度，HPRD50 コーパスで 0.69 程度であることに比べ，低くなっている（0.63 程度）ため仮ラベル付けを行う際にラベルを誤って付与する割合が高くなり，結果として低下を招いていると考えられる．

最後に，特徴づけの際にセマンティックな特徴を用いてデータをサブセットに分け，サブセットごとに与える特徴を変化させて学習を行う *Bui* らの手法 [6] との比較を行う．その結果を Table 13 に示す．

提案手法は *Bui* らの手法と比べ，HPRD50 コーパスにおいて同程度の精度を示した．しかし LLL コーパス, IEPA コーパスにおいては比較的精度が低い結果となった．LLL コーパスにおいては前述のとおり提案手法が精度上昇を図る際に，事例数が少なくうまく働かなかったため *Bui* らの手法に比べ精度が低くなっている．IEPA コーパスに関しては精度に大きな開きが存在した．各学習手法を使用した分類器の精度が低いためで，これの改善には，蛋白質ペア毎の特徴数を増やし，学習精度の底上げを図る必要があると考えられる．一方で提案手法が有用であると確かめられた，データセットが比較的大きい場合，すなわち，HPRD50 コーパスと IEPA コーパスにおいては，Precision の値で *Bui* らの手法を上回った．

5. 結論

本論文では，文献からの PPI の抽出を目的として，複数の分類器を用いた自動抽出の枠組みについて論じた．類似度，信頼度という尺度を利用した半教師あり学習により，訓練データが十分に得られない場合にも効果的に学習が可能となる手法を提案した．

提案手法は，さまざまな学習手法を適用することで複数の分類器を作成するとともに，分類器の類似度，学習方法の信頼度という概念を導入することにより，より信頼できる仮ラベル付与が可能であることを特徴とする．

相互作用情報抽出の評価実験の結果として，データセットがある程度大きな場合に分類器の類似度，学習方法の信頼度を用いた手法において，それを用いない場合よりも良い精度となることを確認した．また，従来手法との比較を行った結果として，同等の精度かそれよりも劣る精度となったが，適合率という観点でみると，従来手法よりも良い結果を示すことが確認された．

しかしながら，未だ精度において十分とは言えないことから，手法そのものに関する改良の余地があると考えられる．今後の課題として，今回は 8 種類の学習手法に固定して実験を行ったが，実験に利用した学習手法以外にも多数の手法が存在するため，その中から学習手法を自動選択することにより信頼性のあるラベル付けを行っていくことが考えられる．また，分類に用いる分類器群の数の増加や，分類器の類似度と学習手法の信頼度の用い方を変えることでも，より信頼性のあるラベル付けが行えるものと考えられ，今後，検討を進めていく予定である．

参考文献

[1] 阿久津 達也, バイオインフォマティクスの数理とアルゴリズム, 共立出版, pp. 187–189 (2007).
[2] 関根 聡, テキストからの情報抽出, 情報処理, Vol. 40, No. 4, pp. 370–373 (1999).
[3] S. Abney, *Semisupervised Learning for Computational Linguistics*, Chapman & Hall/CRC (2007).
[4] X. Zhu and A. B. Coldberg, *Introduction to Semi-Supervised Learning*, Morgan & Claypool, pp. 36–37 (2009).
[5] R.Chowdhary, J.Zhang and J.S.Liu, *Bayesian inference of protein-protein interactions from biological literature*, Bioinfomatics, Vol.25, Issue. 12, pp. 1536–1542 (2009).
[6] Q. C. Bui, S. Katrenko and P. M. A. Sloot, *A hybrid approach to extract protein-protein interactions*, Bioinfomatics, Vol. 27, Issue. 2, pp. 147–265 (2011).
[7] The Stanford Natural Language Processing Group, *Stanford Parser*, 入手先 ⟨http://nlp.stanford.edu/software/lex-parser.shtml⟩ (2011).
[8] S. Pyysalo *et al*, *Protein-protein interaction corpora*, 入手先 ⟨http://mars.cs.utu.fi/PPICorpora/GraphKernel.html⟩
[9] Machine Learning Group at University of Waikato, *Weka 3 :Data Mining Software in Java*, 入手先 ⟨http://www.cs.waikato.ac.nz/ml/weka/⟩
[10] N. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer (1995).
[11] R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers (1993).
[12] J. J. Rodriguez, L. I. Kuncheva and C. J. Alonso, *Rotation Forest:A New Classifier Ensemble Method*, IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 28, No. 10 (2006).

[13] J. G. Cleary and L. E. Trigg, *K*: An Instance-based Learner Using an Entropic Distance Measure*, Proceedings of the 12th International Conference on Machine learning, pp. 108–114 (1995).
[14] L. Breiman, *Random Forests*, Machine Learning Vol. 45, No. 1, pp. 5–32 (2001).
[15] L. Breiman, J. H. Friedman, R. A. Olshen and C. J. Stone, *Classification and Regression Trees*, Wadsworth International Group, Belmont, California (1984).
[16] P. Melville and R. J. Mooney, *Constructing diverse classifier ensembles using artificial training examples*, Proceedings of the Seventeeth International Joint Conference on Artificial Intelligence, pp. 505–510 (2003).
[17] Y. Freund and R. E. Schapire, *Experiments with a new boosting algorithm*, Proc International Conference on Machine Learning, pp. 148–156 (1996).