Vol.2010-MPS-81 No.25
Vol.2010-BIO-23 No.25
2010/12/17

ユーザフィードバックに基づく訓練データ拡張を伴う蛋白質機能情報文抽出に関する研究

宮西一徳[†1]　尾崎知伸[†2]　大川剛直[†3]

蛋白質の機能は，蛋白質の構造解析実験の結果得られ，論文などの文献中に記述され蓄積されている．この機能情報は，新薬の開発や生命現象の解明に必要な情報であるためデータベース化が求められている．そこで，蓄積された大量の文献から機能情報を抽出する手法を提案する．本論文では，文献からの機能情報の抽出を，文献中の各文について機能情報を含むか否かの分類問題として扱う．このような分類問題に対する典型的な機械学習を用いた手法では，あらかじめ与えられた訓練データを用いて分類器を学習するが，十分な訓練データが与えられない場合，高い精度は期待できない．そこで，訓練データを拡張することによって分類精度の向上を図る．訓練データを拡張するため，機能情報文であるかどうかが判明していない文から構成される参照データを用い，ユーザからのフィードバックを基に距離学習を行うことで参照データから精度向上に効果的な文を選択し，訓練データに追加する．評価実験では，少ないフィードバックで精度の向上が見られ，提案手法によりユーザへの負担を軽減しつつ精度向上が実現できることを確認した．

A Method of Extracting Sentences Containing Protein Function Information with Training Data Extension based on User's Feedback

Kazunori Miyanishi,[†1] Tomonobu Ozaki[†2]
and Takenao Ohkawa[†3]

Protein function is clarified by protein structure analysis and the obtained knowledge has been stated in a number of documents. It is expected to construct the database of the function information, because the function information is useful for various application fields such as drug discovery, understanding of life phenomenon, and so on. Then, we propose the method of extracting the function information from a number of documents. In this paper, extraction of protein information is considered as a classification problem, namely, whether each sentence from the target document includes the function information or not is determined. Typically, in the case of addressing such a classification problem, a classifier is learned using the training data previously given. However, the accuracy is not high when the training data is not large enough. Thus, we attempt to improve the accuracy of classification by extending the training data. Effective sentences for getting high accuracy are selected from the reference data aside from the training data set based on user's feedback, and added to the training data. In the experiment, the accuracy is improved by less feedback. Thus, it is confirmed that the training data is appropriately extended based on user's feedback by the proposed method with user's load reduced.

1. はじめに

蛋白質は，他の化合物と相互作用することによって様々な機能を発現することが知られており，生命活動において重要な役割を果たしている[1)]．この蛋白質の機能に関する情報は，蛋白質の構造解析実験の結果として得られ，大量の文献中に記述され公開されている．機能情報は，新薬の開発や生命現象の解明など様々な分野において有用であり，これを容易に利用可能なものとするため，機能情報のデータベース化が求められている．蛋白質に関連するデータベースが多数構築されているが，これらのデータベースに登録されていない有用な情報が，未だに非常に多くの文献中に含まれている．

近年，生物医学文献から重要な情報を抽出する研究が数多く行われている．例えば Tsai ら[2)] や Sun ら[3)] は，単語の綴り上の特徴や接続関係などに基づき Conditional Random Fields(CRF)[4)] を用いた蛋白質名など固有表現抽出のためのアプローチを提案している．また，蛋白質相互作用情報の抽出に関する研究も行われており，Bunescu ら[5)] は様々な情報抽出手法を利用することでヒトの蛋白質名と蛋白質相互作用情報を抽出する手法を提案している．

これらの研究が固有表現や蛋白質相互作用情報の抽出を目的としているのに対して，本研究では蛋白質構造解析に関する文献を対象とし，蛋白質機能情報を抽出するに際してユーザを支援する手法を提案する．提案手法において，蛋白質機能情報の抽出は，文献中の各

†1 神戸大学大学院自然科学研究科
Graduate School of Science and Technology, Kobe University
†2 大阪大学サイバーメディアセンター
Cybermedia Center, Osaka University
†3 神戸大学大学院システム情報学研究科
Graduate School of System Informatics, Kobe University

Vol.2010-MPS-81 No.25
Vol.2010-BIO-23 No.25
2010/12/17

文が機能情報を含むか否かという分類問題として扱う．各文を 1 つの事例とし，キーワードやパターンなどの特徴に照合するかどうかを属性として与え，分類器を学習する．ここでは分類器として Support Vector Machine(SVM)[6] を使用する．分類器の学習において，訓練データが十分ではない場合，高い分類精度が得られない．そこで，現在の分類結果をユーザに提示し，フィードバックを得ることによって訓練データを拡張し，再度学習する処理を繰り返すことで精度向上を図る．このとき，ユーザの負担を軽減するため，精度向上に効果的な文のみを選択し，訓練データを拡張する．具体的には，少ない訓練データに加えて，機能情報を含むか否かのラベル付けがされていないデータ（参照データ）を使用する．参照データ中の文に仮のラベルを付与し，この仮ラベルに対する信頼度を求める．この信頼度とユーザフィードバックに基づく 2 つの訓練データ拡張の手法を提案する．1 つ目は，ユーザからフィードバックされた真のラベルを利用する方法である．2 つ目は，信頼度の高い仮ラベルが付与された文を訓練データに追加する方法である．

評価実験では，フィードバックの回数を少なくした場合，1 回のフィードバックに用いる文の数を少なくした場合のそれぞれの分類精度の推移から提案手法の有効性を評価する．

2. 文の属性

提案手法では，機能情報の抽出を文の分類として扱うため，訓練データを基に学習した分類器を新たな文献に適用し，文献中の各文について判別を行う．分類器は以下の属性を用いて学習する．

- 原子間距離
  文中に残基名や原子名が含まれる場合，三次元空間上での物理的特徴が機能情報を含むかどうかの手掛かりとなり得る．具体的には，残基が他の物質と相互作用する時，残基中の原子は相互作用する物質に接近する．従って，ある文中に記述されている残基と他の物質との三次元空間上での距離が一定の閾値以下の場合，残基はその物質と相互作用していると考えられるため，この文に対して属性値として “1” を付与する．これ以外の場合は “0” を付与する．
- キーワード
  機能情報に関して記述された文に頻出する単語は，分類の手掛かりとなるためキーワードとする．このキーワード（例，“interact”, “bind”, “hydrogen bond” など）を含む文には属性値として “1” を付与し，含まない場合には “0” を付与する．
- パターン
  機能情報を含む文に頻出するパターンも分類に有用であると考えられる．例えば，“<residue> (.)* play (.)* <function>”, “<protein> (.)* contain (.)* <residue>”（<residue> は “Arg21” や “His23” のような残基名，<function>, <protein> はそれぞれ機能，蛋白質の名前を意味する）のようなパターンを定義する．各パターンごとに，照合する場合には属性値として “1” を付与し，照合しない場合には “0” を付与する．

3. 訓練データの拡張

3.1 訓練データ拡張の概要

事前に十分な訓練データが与えられない場合，訓練データのみで学習した分類器では高い精度が期待できない．そこで，ユーザからのフィードバックに基づく訓練データ拡張により精度向上を図る．提案する手法の概要を**図 1** に示す． 初めに訓練データのみで学習した分類器を参照データに適用し，各文に仮のラベルを付与する．ここで，仮のラベルが正しいことの度合いを表す信頼度を導入し，参照データ中の全ての文の信頼度を求める．高い信頼度の仮ラベルが付与された文は訓練データに追加する（拡張手法 I）．信頼度の低い仮ラベルが付与された文はユーザに提示し，フィードバックとして真のラベルを返してもらう．これらの文については，フィードバックされた真のラベルを付与し，訓練データに追加する（拡張手法 II）．以上の処理を繰り返し行うことで，徐々に訓練データを拡張し精度向上を図る．この際，拡張手法 I によって訓練データに追加された文は仮ラベルが付与されているため，次回の繰り返しステップでは参照データに戻し，仮のラベルを付与しなおした上で信頼度を再計算する．この枠組みでは，分類精度の向上に有効な文を選択する必要がある．効果的な文を適切に選択できれば，少ないフィードバック回数で精度向上が実現でき，結果としてユーザの負担を軽減することができる．

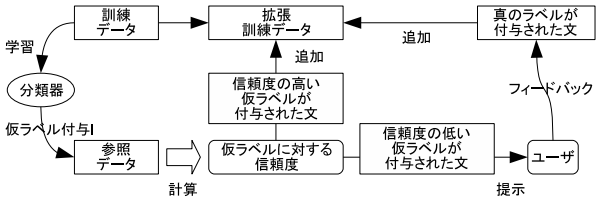


図 1 提案手法の概要

Vol.2010-MPS-81 No.25
Vol.2010-BIO-23 No.25
2010/12/17

3.2 訓練データの拡張

3.2.1 仮ラベルの信頼度

同じラベルが付与された事例は同様の属性を持つと考えられるため，このような事例は属性空間上で近接する．ある文（対象文と呼ぶ）に対して仮のラベルが付与された場合，属性空間上での対象文の周辺に真のラベルで同じラベル値が付与された文が多く存在するとき，対象文の仮ラベルが正しい可能性が高いと考えられる．逆に，周辺に真のラベルで対象文とは異なるラベル値が付与された文が存在する場合，対象文の仮ラベルは誤っている可能性が高い．従って，仮ラベルの信頼度は，属性空間上で周辺の真のラベルが付与された事例（文）の分布を基に計算する．

信頼度の精度を高めるため，属性空間上での真のラベルが付与された文の分布から，真のラベルで同じラベル値が付与された文が近づくように文間の距離を再定義する．半教師つきクラスタリング[7] の分野では，データ間の制約を使用し事例間の距離を学習することによって，より適切なクラスタを得ようとする “Distance Metric Learning”[8),9)] と呼ばれるアプローチがある．Xing ら[8)] のアプローチでは，入力空間 $\mathbb{R}^n$ 上のデータセットを $\{x_i\}_{i=1}^m$ と表し，データ x と y の間の距離 $d(x,y)$ を以下のように定義する．

$$d(x,y) = \| x-y \|_A = \sqrt{(x-y)^T A(x-y)} \tag{1}$$

ここで，A は重みを表す対角行列である．データ x_i と x_j の間での制約 S と D は以下のように与えられる．

$S : (x_i, x_j) \in S$ *if* x_i *and* x_j *are similar*

$D : (x_i, x_j) \in D$ *if* x_i *and* x_j *are dissimilar*

これらの制約の下で，式 (1) の A は以下の最適化問題を解くことによって求められる．

$$\min_A \sum_{(x_i,x_j)\in S} \| x_i - x_j \|_A^2$$

$$s.t. \sum_{(x_i,x_j)\in D} \| x_i - x_j \|_A \geq 1,\ A \succeq O.$$

提案手法では，各文を 1 つの事例とし，文のラベル値をデータ間の制約として扱い，式 (1) によって文間の距離を再定義する．距離を再定義することによって，同じラベル値を持つ文同士が近づき，異なるラベル値を持つ文同士は遠ざかる．従って，距離が再定義された空間上において，ある仮ラベルが付与された文の周辺の真のラベルが付与された文の分布から，その仮ラベルが正しい度合い（信頼度）が適切に求められる．

仮ラベルの信頼度は，再定義された距離空間における真のラベルが付与された文との距離によって求める．近接する事例は同じラベル値を持つ傾向があると考えられる．仮ラベルが付与された文について，同じラベル値の真のラベルを持つ文が多く近接している場合，この仮ラベルは正しい可能性が高い．一方，異なるラベル値の真のラベルを持つ文が近隣に数多く存在する場合，仮ラベルが誤っている可能性が高いと考えられる．この観点から，ある仮ラベルについて同じラベル値の真のラベルを持つ文との距離に基づく信頼度を類似信頼度 f と定義する．また，異なるラベル値の真のラベルを持つ文との距離に基づく信頼度を非類似信頼度 g と定義する．

定義 1（類似信頼度と非類似信頼度）

s_x は仮ラベル l_x が付与された文とし，s_1, s_2, ..., s_n は真のラベル l_1, l_2, ..., l_n が付与された文とする．s_x と s_i の距離を $d_i(s_x, s_i)$ と表すとき，類似信頼度 $f(s_x)$ と非類似信頼度 $g(s_x)$ を以下のように定義する．

$$f(s_x) = \sum_i \frac{1}{d_{(s_x,s_i)}}\ (l_x = l_i) \tag{2}$$

$$g(s_x) = \sum_i \frac{1}{d_{(s_x,s_i)}}\ (l_x \neq l_i) \tag{3}$$

■

類似信頼度が高く，非類似信頼度が低い文ほど仮ラベルが正しいと考えられるため，最終的な信頼度 r を以下のように定義する．

定義 2（信頼度）

$$r(s_x) = f(s_x) - g(s_x) \tag{4}$$

■

3.2.2 訓練データ拡張手法

信頼度に基づいて，参照データから文を選択し訓練データを拡張する手続きを**図 2** に示す．ここで，T_t と T_b はそれぞれ拡張手法 I と II の閾値である．T_t は，信頼度の高い仮ラベルが付与され訓練データに追加される文の数に関する閾値である．T_b は，付与された仮ラベルの信頼度が低く，ユーザに提示することにより，真のラベルを問う必要がある文の数に関する閾値である．信頼度の高い真のラベルは正しい可能性が高いため，このようなラベルを持つ文は訓練データに追加する（拡張手法 I）．一方，信頼度の低い仮ラベルを持つ文

Vol.2010-MPS-81 No.25
Vol.2010-BIO-23 No.25
2010/12/17

Procedure：参照データからの文選択とフィードバックに基づく訓練データ拡張
参照データ中の文 $\{S_j\}_{j=1}^m$ について：
for $j = 1...m$
　式 (4) に基づき $r(S_j)$ を計算
end
From $\{r(S_j)_{j=1}^m\}$,
上位 $T_t \rightarrow$ を訓練データに追加 (拡張手法 I)
下位 $T_b \rightarrow$ をユーザに提示 (拡張手法 II)

図 2　文選択とフィードバックによる訓練データ拡張の手順

は，ユーザフィードバックによって真のラベルを付与し，訓練データに追加する（拡張手法 II）．仮ラベルに対する信頼度は全て，新たなフィードバックによって更新されるため，拡張手法 I によって訓練データに追加された文は，次回の繰り返しステップでは参照データに戻し，再び T_t 個の文を選択し直す．従って，訓練データの文の数は，拡張手法 II によって真のラベルがフィードバックされ訓練データに追加される T_b だけステップごとに増加する．

4. 評　　価

表 1 に示す文献を使用し，提案手法の有効性を評価する．各文献は PDB(Protein Data Bank)[10] から参照されており，PDB-ID は PDB 内に登録されている蛋白質に対する識別子である． これらの文献中の固有表現（蛋白質名や残基名，原子名など）はあらかじめタグ付けされているものとする．実験では，表 1 に示す 23 文献のうちランダムに選択した 7 つの文献を訓練に使用した．そのうち 10 個の文を真のラベルが付与された初期訓練データとして使用し，残りの文を参照データとして使用した．さらに，訓練に使用しなかった 16 の文献を評価用として用いた．異なる 8 通りの文献の組み合わせで実験を行った．

SVM を分類器として使用し，1 つの文を 1 つの事例として扱う．2 で示した属性を使用し訓練を行う．属性として使用したキーワードの数は 45，パターンの数は 19 である．

閾値 T_t は 25，T_b は 6 から 200 の間で変化させて実験を行った．つまり，信頼度の高い仮ラベルを持つ 25 個の文を，各繰り返しステップにて訓練データに追加する．1 回のフィードバックで返される文の数を 6 から 200 の間で変化させた時の精度の推移を**図 3** に示す．横軸はフィードバックによって返される真のラベルを付与した文の数を示しており，フィードバックが返されるたびに増加していく．精度は，評価用文献に対する分類結果の評価尺度である F 値により評価した． 1 回のフィードバックで返される文の数が 30(つまり $T_b = 30$) の時，早い段階で精度が高くなっている．例えば，30 文ずつフィードバックした場合，フィードバックによって 60 個の文に真のラベルが付与された時点での F 値は約 62%である．これに対して，60 文ずつフィードバックした場合，2 回のステップつまり 120 個の文に真のラベルを付与した段階でほぼ同じ F 値に達している．結果として，30 文ずつフィードバックした方が 60 文ずつフィードバックするよりも 60 文少ない段階で同じ精度に達している．

表 1　実験で用いた文献データ

PDB-ID	文献中の文数	機能情報文の数	PDB-ID	文献中の文数	機能情報文の数
1a0f	382	46	1a0h	359	26
1a0k	683	19	1a0o	148	12
1a0q	295	23	1a1s	285	24
1a23	528	5	1a26	243	13
1a3a	544	17	1a3h	275	8
1a3l	272	23	1a3r	299	21
1a3s	306	7	1a3y	209	3
1a4j	190	13	1a5a	113	10
1a5h	296	39	1a5i	324	73
1a5v	277	20	1a5y	291	33
1a5z	428	8	2a2g	365	13
2a39	312	4			

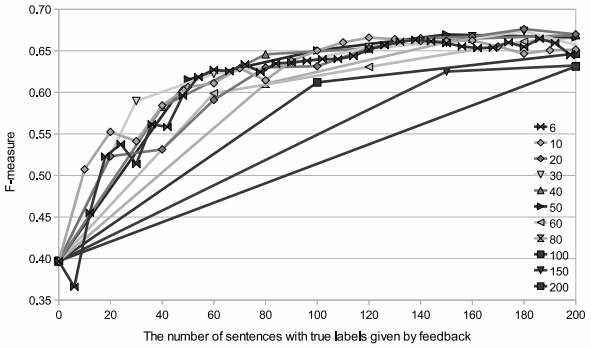


図 3　提案手法の精度 ($T_b = 6 \sim 200$)

Vol.2010-MPS-81 No.25
Vol.2010-BIO-23 No.25
2010/12/17

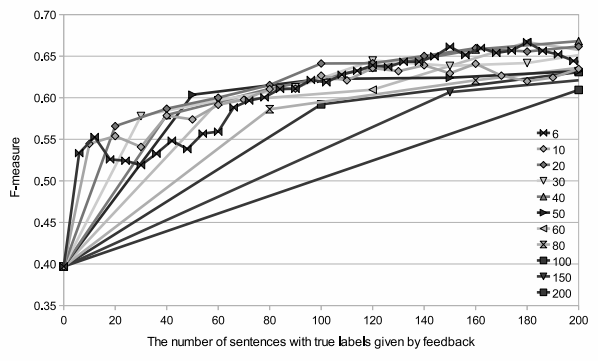
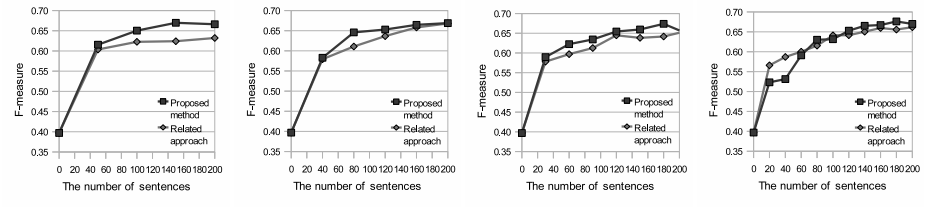


図 4 関連手法の精度 ($T_b = 6 \sim 200$)

従って，T_b が 30 以上の時つまり 1 回のフィードバックが 30 文以上の時は，少しずつフィードバックする方が精度が高くなることが分かる．しかしながら，T_B が 30 より小さい場合，初期段階での精度が低い．これは，正しく学習するための真のラベルを持つ文の数が少ないためと考えられる．特に T_b が 6 の時，初めのステップで精度が低下し，2 回目以降で上昇に転じ，60 文前後で安定している．

提案手法と同様に，選択した事例をユーザに提示し正解をフィードバックしてもらうことで精度の向上を図る枠組みは，一般に能動学習と呼ばれている．能動学習の分野での関連研究として，ユーザに提示する文選択の基準として SVM の超平面からの距離を使用するアプローチがある[11)–13)]．超平面に近い事例の信頼度は低いという考えに基づき，そのような事例をユーザに提示し，正しいラベルを付与するというアプローチである．比較実験として，このアプローチを適用し，仮ラベルに対する信頼度を超平面からの距離に基づいて計算する．精度の推移を**図 4** に示す． 提案手法と比べて，全体的な精度は低くなっている．特に，初期の段階での精度が低い．

この違いを明らかにするため，T_b が 50, 40, 30, 20 の時の精度の比較を**図 5** (a) – (d) に示す． T_b が 50, 40, 30 の時，提案手法の方が精度が高くなっている．$T_b = 20$ の場合，初期の段階では提案手法の方が精度が低いが，真のラベルを持つ文の数が 60 以上でほぼ同じ精度となっている．T_b が一定の値以上であれば，提案手法による精度の方が安定している．一般に，最も適切な T_b を正確かつ事前に求めることは困難であり，より広い範囲の T_b に



(a) $T_b = 50$ (b) $T_b = 40$ (c) $T_b = 30$ (d) $T_b = 20$

図 5 提案手法と関連手法との精度比較

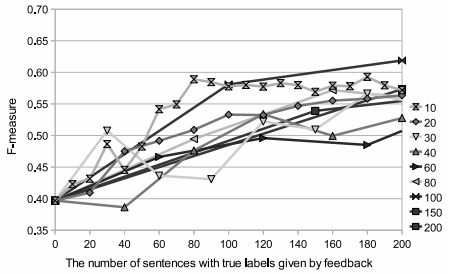


図 6 ランダムに文選択した場合の精度

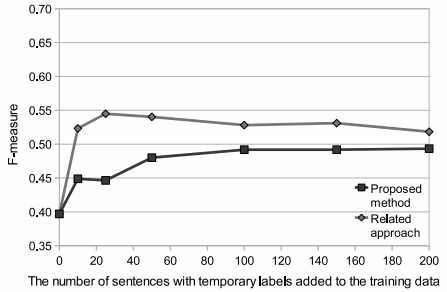


図 7 フィードバック無しの場合の制度 ($T_b = 0$)

おいて精度が高く，さらにより早い段階で収束する提案手法の方が関連手法より優れていると言える．

ユーザに提示する文をランダムに選択した場合の精度の推移を**図 6** に示す． 繰り返しステップにおいて，ランダムに訓練データに追加される文によって精度は悪影響を受ける場合が多く，収束する様子も見られない．このことから，提案手法での文選択の基準が精度向上に効果的であることが分かる．次に，拡張手法 I のみ（ユーザフィードバックが無い場合）の精度と比較するため，訓練データに追加する文の数（T_t）を変化させた時の精度の推移を**図 7** に示す．提案手法，関連手法両方とも低い精度で収束している．特に提案手法ではより低い値となっており，ユーザフィードバックが精度向上に大きく貢献していることが分かる．

以上のことから，フィードバックの回数ならびに 1 回のフィードバックで返される文の数ともに少ない場合に，提案手法での精度向上が見られた．さらに，T_b が一定の値以上の場合，フィードバックの繰り返しに伴い，精度が単調に増加することが観測された．従って，

Vol.2010-MPS-81 No.25
Vol.2010-BIO-23 No.25
2010/12/17

距離学習に基づく信頼度が効果的であることが確認できた．

5. おわりに

本論文では，ユーザフィードバックを使用した訓練データ拡張を伴う蛋白質機能情報文抽出手法を提案した．提案手法では，参照データ中の文に対して仮ラベルを付与し，この仮ラベルの信頼度を求めることによって訓練データを拡張するために適切な文を選択する．この時，ユーザフィードバックを基にして距離学習を行うことによって信頼度を求める．

能動学習の分野での他手法と比較し，フィードバックの繰り返し回数が少ない場合に提案手法の精度が高くなった．さらに，1 回のフィードバックで返される文の数が少ない場合にも，提案手法の精度が上回った．従って，提案手法においてユーザフィードバックにより訓練データが適切に拡張できることが確認できた．さらに，この結果はユーザの負担を軽減するという目的に適っていることを示している．

今後の展望としては，仮ラベルが誤っていたためにフィードバックによって訂正された文に共通する特徴を利用して再学習することで，次回の繰り返しステップでの分類器の精度向上が実現できると考えられる．

参考文献

1) J. M. Berg, J. L. Tymoczko and L. S. Stryer. Biochemistry fifth edition. *WH Freeman and Company.*, 423: 436–437, 2002.
2) R. T. H. Tsai, C. L. Sung, H. J. Dai, H. C. Hung and T. Y. Sung. NERBio: using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC Bioinformatics*, 7 (Suppl 5): S11, 2006.
3) C. Sun, Y. Guan, X. Wang and L. Lin. Biomedical named entities recognition using conditional random fields model. *Fuzzy Systems and Knowledge Discovery*, 1279–1288, Springer Berlin / Heidelberg, 2006.
4) J. Lafferty, A. McCallum and F. Pereira. Conditional random fields: probabilistic models for segmenting and labeling sequence data. *Proceedings of the International Conference on Machine Learning (ICML '01)*, 282–289, 2001.
5) R. Bunescu, R. Ge, R. J. Kate, R. J. Mooney, Y. W. Wong, E. M. Marcotte and A. Ramani. Learning to extract proteins and their interactions from medline abstracts. *Proceedings of the ICML-2003 Workshop on Machine Learning in Bioinformatics*, 46–53, 2003.
6) V. N. Vapnik. The nature of statistical learning theory. *Springer*, 1995.
7) D. Cohn, R. Caruana and A. McCallum. Semi-supervised clustering with user feedback. *Constrained Clustering: Advances in Algorithms, Theory, and Applications*, 17–31, 2008.
8) E. P. Xing, A. Y. Ng, M. I. Jordan and S. Russell. Distance metric learning with application to clustering with side-information. *Advances in neural information processing systems*, 521–528, 2003.
9) A. Bar-Hillel, T. Hertz, N. Shental and D. Weinshall. Learning distance functions using equivalence relations. *Proceedings of the Twentieth International Conference on Machine Learning (ICML)*, 20(1): 11–18, 2003
10) H. M. Berman, T. Battistuz, T. N. Bhat, W. F. Bluhm, P. E. Bourne, K. Burkhardt, Z. Feng, G. L. Gilliland, L. Iype, S. Jain, P. Fagan, J. Marvin, D. Padilla, V. Ravichandran, B. Schneider, N. Thanki, H. Weissig, J. D. Westbrook, C. Zardecki. The protein data bank. *Acta Crystallographica Section D: Biological Crystallography*, 58(6): 899–907, 2002.
11) M. Sassano. An empirical study of active learning with support vector machines for Japanese word segmentation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, 505–512, 2001.
12) G. Schohn and D. Cohn. Less is more: active learning with support vector machines. *Proceedings of the 17th International Conference on Machine Learning*, 839–846, 2000.
13) S. Tong and D. Koller. Support vector machine active learning with applications to text classification. *The Journal of Machine Learning Research*, 2: 45–66, 2002.