2008－BIO－14 (12)
2008／9／19

異種ネットワーク統合によるタンパク質機能予測

加藤 毅[†] 鹿島久嗣[††] 杉山 将[†††]

今日では，分子生物学で用いるデータは非常に巨大化しており，また，データの種類も多種多様である．これまでは，これらのデータを解析するために，しばしばカーネル法と呼ばれる解析法が試みられてきた．しかし，カーネル法は多くのメモリと計算量が必要である．本発表では，小さいメモリで高速に計算するために，データをネットワークであらわし，それらから自動的に取捨選択してクラス分類を行う算法を提案する．提案法をタンパク質の機能を予測する実問題を使った実験結果を通して，従来法と同等な予測性能を持ちながら，30 倍以上高速であることを示す．

Protein function prediction by integration of heterogeneous biological networks

Tsuyoshi Kato,[†] Hisashi Kashima[††] and Masashi Sugiyama[†††]

The size and type of data used in molecular biology are getting enormous these days. Computational analyses of those data have been attempted using kernel methods, but they require a large memory size. In this paper, we represent the data by networks and devise a new algorithm for accurate data analysis. The proposed method allows us to efficiently integrate heterogeneous networks.

1. 導 入

実験技術の向上に伴って，分子生物学で用いられるデータのサイズは日々大きくなっている．これまでは，これらのデータを解析するために，しばしばカーネル法と呼ばれる解析法が試みられてきた．しかし，カーネル法は多くのメモリと計算量が必要である．そこで筆者らは，小さいメモリで高速に計算するために，ネットワークという新たなデータの表現方法に注目している．ネットワークは，通常，無方向グラフで表される．ノードが例題を表し，エッジが例題間の何らかの関係を表す．たとえば，タンパク質相互作用ネットワークの場合，ノードがタンパク質で，エッジがタンパク質間の相互作用となる．遺伝子発現データのネットワークは，ノードは同じく遺伝子，発現データが類似しているときエッジをつなぐことによって構築できる．機能未知のタンパク質の機能をあるタンパク質ネットワークから予測したい場合，タスクはネットワーク上のノードのラベルを識別することになる．ネットワークに含まれる一部のタンパク質の機能が分かっている場合は，トランスダクティブ (transductive)[3] なタスクになる．このような設定における学習を**グラフ型学習**と呼ぶことにする．

グラフ型学習のためによく使われているアプローチは**ラベル伝播法**[4] である．ラベル伝播法では，ネットワーク上で近いノードどうしは同じラベルを持ちやすいという仮定を使っている．

本論文では，複数のデータ，すわなち，複数のネットワークが所与の状況を考える．タンパク質の機能予測の場合，様々なデータを利用することができる．たとえば，遺伝子発現データ，アミノ酸配列，系統発生データ，局在データなどである．ラベル伝播法の場合，それぞれのデータセットをネットワークとして用いる．異なるデータセットは異なる情報を持っていることが多いので，これらの情報を効果的に統合すれば，予測性能を向上できると期待される．

複数のネットワークの統合には，グラフラプラシアンの線形和をとる方法が考えられる．Tsuda ら[3] は自動的に重み係数を決定する方法を提案している．しかし，Tsuda らの方法は，予測に役立たないネットワークほど大きな重み係数を与える算法になっている．したがって，よって，Tsuda らの方法は無関連なネットワークがデータ集合に含まれている状況では頑健性に欠ける (文献[2] 6 節参照)．

本論文では，異種のネットワークを効果的に統合して頑健に予測を行う新しい算法を提案する．Tsuda

† お茶の水女子大学
Ochanomizu University
†† IBM 東京基礎研究所
Tokyo Research Laboratory of IBM Research
††† 東京工業大学
Tokyo Institute of Technology

ら[3] と同様，提案する算法もグラフラプラシアンの線形結合における重み付けを自動的に行う．Tsuda ら[3] の方法に対する提案法の利点は，予測に重要なネットワークほど大きな重みを与え，予測に無関連なネットワークほど小さな重みを与える点があげられる．提案法の確率モデルは頑健な予測を行うために Student-t 分布を用いている．Student-t 分布を含むことから，提案モデルは潜在変数モデルと解釈でき，潜在変数の期待値はグラフラプラシアンの重み係数であることが自然に導かれる．本論文では，紙面の制約から算法のみ示す．確率モデルからの導出は，文献[2] 5 節を参照されたい．提案法は効率的であるという利点もある．解析的にも効率性を保証できる（文献[2] 6 節）が，本論文では，5.3 節で実験的に効率性を示す．また，タンパク質の機能を予測するタスクに適用して，提案法の有効性を示す．

2. グラフ型学習問題

まず，グラフ型学習問題を定義しよう．グラフ型学習問題は次の 2 つの設定を含んでいる：一つは**一つのネットワークから学習**する場合，もう一つは**複数のネットワークから学習**する場合である．解析の対象として n 個のタンパク質が所与としよう．最初の $\ell(<n)$ 個のタンパク質は $y_1,\ldots,y_\ell \in \{\pm 1\}$ でクラスラベルが与えられているとし，残りの $(n-\ell)$ 個のタンパク質のクラスラベルは未知とする．本論文で考えるタスクはこれら $(n-\ell)$ 個のタンパク質のクラスラベルを予測することである．一つのネットワークから学習する場合，n 個のノードの無方向ネットワーク 1 個を用いる．そのネットワークは，正規化された対称な隣接行列 $\boldsymbol{A} \in \mathbb{S}^n$ で表現される．すなわち，$\boldsymbol{A}$ は $\boldsymbol{A} = \boldsymbol{A}^\top$，$\boldsymbol{A}\boldsymbol{1}_n = \boldsymbol{1}_n$，$A_{ii} = 0$ $(\forall i)$ を満たし，全要素が非負である．エッジの集合は $\mathcal{E} \equiv \{(i,j) \,|\, A_{ij} > 0,\ i,j = 1,\ldots,n\}$. で表される．これが一つのネットワークから学習する場合の問題設定である．本論文では複数のネットワークから学習する場合を考える．ネットワークの個数が K のとき，それぞれに対応する K 個の隣接行列 $\boldsymbol{A}_k$ $(k = 1,\ldots,K)$ が所与とするものである．

3. 既存のラベル伝播法

本節では，1 つのネットワークからグラフ型学習を行う方法[4] を紹介する．ここでは，ネットワーク $\boldsymbol{A}$ およびクラスラベル $\boldsymbol{y}$ からスコアベクトル $\boldsymbol{f} \in \mathbb{R}^n$ を決定する．スコアベクトル $\boldsymbol{f} \in \mathbb{R}^n$ を計算するための典型的な算法は，正則化最小二乗問題

$$\hat{\boldsymbol{f}} \equiv \operatorname*{argmin}_{\boldsymbol{f}} \left(\beta_{\mathrm{y}} \sum_{i=1}^{\ell} (y_i - f_i)^2 + \beta_{\mathrm{bias}} \sum_{i=1}^{n} f_i^2 + \beta_{\mathrm{net}} \sum_{(i,j)\in\mathcal{E}} A_{ij} (f_i - f_j)^2 \right)$$

で定式化されている．ただし，β_{y}, β_{bias} および β_{net} は定数である．もしあるノードのスコアが閾値以上ならば，そのノードは陽性と識別し，それ以外は陰性と識別する．この最小化問題は，$\boldsymbol{L} \equiv \mathrm{diag}\,(\boldsymbol{A}\boldsymbol{1}_n) - \boldsymbol{A}$, で定義されるグラフラプラシアンを使うと，

$$\min_{\boldsymbol{f}} \beta_{\mathrm{y}} (\boldsymbol{f} - \boldsymbol{y})^\top \boldsymbol{G} (\boldsymbol{f} - \boldsymbol{y}) + \boldsymbol{f}^\top (\beta_{\mathrm{bias}} \boldsymbol{I}_n + \beta_{\mathrm{net}} \boldsymbol{L}) \boldsymbol{f} \tag{1}$$

と表せる．ただし，$\boldsymbol{G} \in \mathbb{S}_+^n$ は対角行列で，第 1 対角要素から第 ℓ 対角要素まで 1，それ以外は 0 とするものである．つまり，$\boldsymbol{G}$ の第 i 対角要素の値は，第 i ノードのクラスラベルが所与の場合は 1 で，それ以外は 0 である．ここで $\boldsymbol{y}$ は n 次元ベクトルで

$$\boldsymbol{y} \equiv [y_1,\ldots,y_\ell,0,\ldots,0]^\top \in \mathbb{R}^n.$$

と定義した．

4. 提案するラベル伝播法

この節では，複数のネットワークから頑健にラベル伝播を行う新しい算法を提案する．このタスクは K 個のネットワーク $\{\boldsymbol{A}_k\}_{k=1}^K$ およびクラスラベル $\boldsymbol{y}$ からスコアベクトル $\boldsymbol{f} \in \mathbb{R}^n$ を予測するものである．このタスクに対する一つの選択は，まず複数のネットワークを一つに統合してから前節で述べたラベル伝播法に適用する方法である．ネットワークを統合する方法の最も単純な方法の一つに，重ね合わせが考えられる．本研究では，有益な情報を強調するために，K 個のネットワークの重みつき線形結合による統合を考える．重み係数を $\bar{\boldsymbol{u}} \in \mathbb{R}_+^K$ で表すことにする．すると，統合された隣接行列とグラフラプラシアンは，それぞれ，$\boldsymbol{A}_{\mathrm{int}}(\bar{\boldsymbol{u}}) = \sum_{k=1}^K \bar{u}_k \boldsymbol{A}_k$ および，$\boldsymbol{L}_{\mathrm{int}}(\bar{\boldsymbol{u}}) = \sum_{k=1}^K \bar{u}_k \boldsymbol{L}_k$. で与えられる．いったん重み係数を決定できれば，前節で述べた標準的なラベル伝播法

$$\begin{aligned}\hat{\boldsymbol{f}} &= \operatorname*{argmin}_{\boldsymbol{f}} \Big(\beta_{\mathrm{y}} (\boldsymbol{f} - \boldsymbol{y})^\top \boldsymbol{G} (\boldsymbol{f} - \boldsymbol{y}) \\ &\qquad + \boldsymbol{f}^\top (\beta_{\mathrm{bias}} \boldsymbol{I}_n + \beta_{\mathrm{net}} \boldsymbol{L}_{\mathrm{int}}(\bar{\boldsymbol{u}})) \boldsymbol{f} \Big) \\ &= \left(\boldsymbol{G} + \frac{\beta_{\mathrm{bias}}}{\beta_{\mathrm{y}}} \boldsymbol{I}_n + \frac{\beta_{\mathrm{net}}}{\beta_{\mathrm{y}}} \boldsymbol{L}_{\mathrm{int}}(\bar{\boldsymbol{u}}) \right)^{-1} \boldsymbol{G}\boldsymbol{y}. \end{aligned} \tag{2}$$

でスコアを決定できる．では，どのように重み係数を計算するか？提案法では，現在のスコア $\boldsymbol{f}$ を使って，次の更新則で反復的に重み係数を決定する：

$$\bar{u}_k = \frac{\nu + n}{\nu + \beta_{\mathrm{net}} \boldsymbol{f}^\top \boldsymbol{L}_k \boldsymbol{f}}. \tag{3}$$

ただし，ν は正の定数である．この更新則は EM 算法を使って自然に導かれる (文献[2] 4 節参照)．ここで

$$\boldsymbol{f}^\top \boldsymbol{L}_k \boldsymbol{f} = \sum_{(i,j)\in\mathcal{E}_k} A_{ij}^{(k)} (f_i - f_j)^2,$$

を満たすことに注意されたい．ただし，$A_{ij}^{(k)}$ は $\boldsymbol{A}_k$ の第 (i,j) 要素である．もし，各ノードが隣接するノードと同じクラスラベルを持つ傾向にあるならば，ラベル伝播法はうまく働く．言い換えると，ラベル伝播法にとって有益なネットワークは隣接するノードは同じ

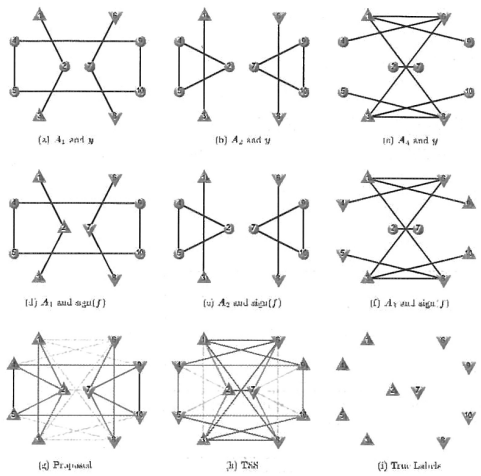

図 1 人工データによるデモ．(a),(b),(c) にこのデモで用いた 3 つのネットワーク $\boldsymbol{A}_1$, $\boldsymbol{A}_2$ および $\boldsymbol{A}_3$ が所与とする．上向き三角形，下向き三角形，および円は，それぞれ $y_i = +1$, $y_i = -1$ を持つノード，およびラベルなしノードを示している．真のクラスラベルは (i) に示す．(d),(e),(f) は，個々のネットワークでラベル伝播させた予測結果である．提案法と TSS の結果を (g),(h) に示す．エッジの濃淡はネットワークの重み係数を表している．

予測値を得やすいという性質を持っていることになる．このような理由から，$\boldsymbol{f}^\top \boldsymbol{L}_k \boldsymbol{f}$ の値は，その予測タスクに対して無関係なネットワークほど大きな値になり，有益なネットワークほど $\boldsymbol{f}^\top \boldsymbol{L}_k \boldsymbol{f}$ の値は小さくなる．式 (3) の分母に $\boldsymbol{f}^\top \boldsymbol{L}_k \boldsymbol{f}$ の項があるので，有益なネットワークほど重み係数は大きくなり，無関係なネットワークほど重み係数は小さくなる．提案算法は次のようにまとめられる：

1: 重み係数の初期値を $\bar{\boldsymbol{u}} = \boldsymbol{1}_K / K$ と設定する;
2: **repeat**
3: 式 (2) を使ってスコア $\boldsymbol{f}$ を更新する;
4: 式 (3) を使って重み係数 $\bar{\boldsymbol{u}}$ を更新する;
5: **until** 収束.

5. 実　　験

5.1 人工データ

図 1(a),(b),(c) は，それぞれ，このデモで用いる 3 つのネットワークを用いる．これら 3 つのネットワークは，10 個の共通のノードを持っている．10 個中，最初の 5 個のノードは正例で，残りの 5 個のノードは負例である．第 1,3,6,8 ノードはクラスラベルが所与とする．$\boldsymbol{A}_1$ および $\boldsymbol{A}_2$ に含まれる多くのエッジは同じクラスラベルのノードを繋いでいる．一方，$\boldsymbol{A}_3$ に含まれる多くのエッジははクラスラベルが異なるノードを繋いでいる．したがって，$\boldsymbol{A}_1$ と $\boldsymbol{A}_2$ はそのタスクにとって有益であり，逆に $\boldsymbol{A}_3$ は無関係である．

図 1(d),(e),(f) は，それぞれ一つのネットワークを既存の方法を使って個々にラベル伝播法を適用した結果である．$\boldsymbol{A}_1$ では，第 2,7 ノードのクラスラベルは正しく予測された．しかし，第 4,5,9,10 ノードは予測されなかった．なぜなら，これら 4 つのノードはどのラベルありノードとも連結していないからである．$\boldsymbol{A}_2$ も，同じ理由でどのノードも予測されなかった．$\boldsymbol{A}_3$ では，第 4,5,9,10 ノードの予測が得られたが，その予測は誤りであった．これは，$\boldsymbol{A}_3$ に含まれる無関係なエッジのためである．この結果より，それぞれのネットワークから個々にラベル伝播法を用いたのでは正しい予測は得られない．よって，よりよい予測を得るには，これらのネットワークを統合する必要がある．

では，提案法を使って統合したネットワークからラベル伝播した結果を示そう．図 1(g) は提案法の予測結果である．すべてのノードのクラスラベルが正しく予測されている．得られた重みは $\bar{\boldsymbol{u}} = [2.98, 3.78, 0.53]^\top$ である．図中のエッジの濃淡は重みの値 $\bar{u}_k$ を表している．3 番目のネットワークの重み $\bar{u}_3$ は自動的に小さな重み係数が与えられている．その結果，すべてのノードのクラスラベルは正しく予測された．提案法と同様，Tsuda ら[3] も線形結合の重み係数を自動的に与える算法を提案している．この算法を TSS と略記する．図 1(h) は TSS の予測結果である．TSS は $\bar{\boldsymbol{u}} = [4.00, 2.00, 4.00]^\top$ という重み係数を与えた．このように，TSS は 3 番目のネットワークに大きな重みを与えた．これによって不十分な予測となってしまった．

5.2 Protein Function Prediction

次にタンパク質の機能を予測する実験を行った．具体的には，各タンパク質がリボソーマルか否か判定する 2 クラス識別問題を行った．用いたデータセットは全部で 760 のタンパク質を含み，そのうち 92 個がリボソーマルである．ネットワークとしては，タンパク質相互作用ネットワーク $\boldsymbol{A}_{\mathrm{vm}}$ と遺伝子発現ネットワーク $\boldsymbol{A}_{\mathrm{e}}$ を用いた．このほかに，不要なデータを除去できる能力を確認するため，ノード番号を無作為に入れ替えることによって作成した囮ネットワーク $\boldsymbol{A}_{\mathrm{r1}}$ および $\boldsymbol{A}_{\mathrm{r2}}$ も用いた．実験条件の詳細は文献[2] を参照されたい．

提案法は既存の 2 つの方法と比較した：比較対象の一つは TSS[3] で，もう一つは SDP/SVM[1] である．TSS は，提案法と同じく，ネットワークの線形結合を得るが，SDP/SVM は，ネットワークをカーネル化したものを統合するので，SDP/SVM から統合したネットワークを得ることはできない．提案法，および比較手法のハイパーパラメータの値は訓練データセットに交差確認法を適用することにより決定した．表 1 に ROC カーブの AUC (ROC スコア) で評価した結果を示す．各行は，ROC スコアによる予測性能と，それぞれのネットワークに対する重み係数 $\bar{u}_{\mathrm{vm}}, \bar{u}_{\mathrm{e}}, \bar{u}_{\mathrm{r1}}$，お

表 1 タンパク質機能予測の結果．左の表は，$\boldsymbol{A}_{\mathrm{vm}}$ と $\boldsymbol{A}_{\mathrm{e}}$ の2つだけを使ったときの，ROC スコア (ROC カーブの AUC)，および重み係数を報告する．右の表は，2つの囮ネットワーク $\boldsymbol{A}_{\mathrm{r1}}$，$\boldsymbol{A}_{\mathrm{r2}}$ を加えたときの結果である．重みは合計が 1 になるように正規化している．

Method	ROC score	$\bar{u}_{\mathrm{vm}}$	$\bar{u}_{\mathrm{e}}$
Proposed	1.000	0.377	0.623
TSS	0.999	0.500	0.500
SDP/SVM	0.999	-	-

Method	ROC score	$\bar{u}_{\mathrm{vm}}$	$\bar{u}_{\mathrm{e}}$	$\bar{u}_{\mathrm{r1}}$	$\bar{u}_{\mathrm{r2}}$
Proposed	0.998	0.279	0.408	0.16	0.154
TSS	0.721	0.200	0.000	0.400	0.400
SDP/SVM	0.999	-	-	-	-

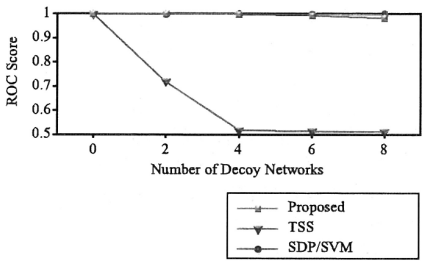

図 2 囮ネットワークの個数に対する予測精度．

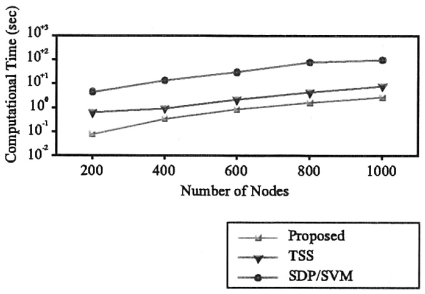

図 3 計算時間．

および $\bar{u}_{\mathrm{r2}}$ を示している．Wilcoxon 検定によって ROC スコアの差の統計的有意性を検定した．囮ネットワークなしでは，3つの手法の差に統計的有意性は見られなかった．2つの囮ネットワークを加えると，TSS の性能は有意に下がった (P 値 = 0.006)．これは，TSS が囮ネットワークに大きな重みを与えたからである．これに対して，提案法と SDP/SVM は有意な差は検出されなかった (P 値はそれぞれ 0.100 および 0.181)．

さらに，提案法の頑健性を調査するために，囮ネットワークの個数を増やした．図 2 は，提案法，TSS，および SDP/SVM の ROC スコアを示している．TSS は囮ネットワークの個数が増えるにつれ，かなり性能が悪くなった．それに対して，提案法と SDP/SVM は関連のないネットワークの個数が増えても予測性能をほぼ維持することが分かった．

まとめると，提案法は最適に統合されたネットワーク $\boldsymbol{L}_{\mathrm{int}}(\bar{\boldsymbol{u}})$ も供給するだけではなく，SDP/SVM と同等な識別性能を得ることが分かった．

5.3 計算時間

次に各手法の計算時間を解析した．MNIST 数字画像データベースから $\boldsymbol{A}_{\mathrm{n1}},\ldots,\boldsymbol{A}_{\mathrm{n5}}$ の5つのネットワークを構成した (詳細は文献[2] 7.D 節参照)．MNIST データベースから無作為に数字画像を抽出し，様々なノードの個数における計算時間を測定した．5 回試行し，平均計算時間を図 3 をプロットした．SDP/SVM はすべてのケースにおいて提案法より 30 倍以上遅かった．また，提案法は TSS よりも十分に高速であるという結果を得た．この理由として，TSS は MATLAB の関数 `fmincon` を使っている．この関数は，一般にあまり効率的ではないことが知られている勾配ベースの数値算法によって実装されている．

文献[2] 7.E 節では，提案法はさらに計算時間を 1/5 程度に減らせることも示されており，故に，提案法は SDP/SVM と同等な予測精度を有しながら，150 倍以上高速な算法であるといえる．

参 考 文 献

1) Argyriou, A., Herbster, M. and Pontil, M.: Combining Graph Laplacians for Semi–Supervised Learning, *Advances in Neural Information Processing Systems 18* (Weiss, Y., Schölkopf, B. and Platt, J.(eds.)), MIT Press, Cambridge, MA, pp.67–74 (2006).

2) Kato, T., Kashima, H. and Sugiyama, M.: Robust Label Propagation on Multiple Networks, *IEEE Trans. on Neural Networks*. in press.

3) Tsuda, K., Shin, H. and Schölkopf, B.: Fast protein classification with multiple networks, *Bioinformatics*, Vol. 21, No. 2, pp. i59–i65 (2005).

4) Zhu, X., Ghahramani, Z. and Lafferty, J.: Semi-supervised learning using Gaussian fields and harmonic functions., *Proceedings of the Twentieth International Conference on Machine Learning* (Fawcett, T. and Mishra, N.(eds.)), San Francisco, CA, AAAI Press (2003).