機械学習アプローチに基づく生物学データ解析法に関する研究

A study on machine learning approach for biological data analysis

化学研究所バイオインフォマティクスセンター 生命知識工学研究領域　烏山 昌幸

背景と目的

計測技術の発展に伴って DNA 配列情報等をはじめとする生物学上の様々なデータが大量に蓄積されるようになってきた．一方で，膨大なデータからどのように有益な情報を抽出するかが問題となっている．本研究では，特に機械学習と呼ばれる統計学と計算機科学の融合技術に注目する．機械学習を用いて生物学上のデータを扱う上で重要となる点のひとつは構造を持つデータをどう扱うかということになる．古典的な数値や文字列の集合データとは異なり，データが固有の内部構造，関係性を持っている場合が多い．典型的な例としてはタンパク質の相互作用ネットワークの場合，各タンパク質間で物理的な相互作用を起こすペアをエッジでつないだグラフとしてデータが表現され，各ノード（タンパク質）の性質を推定することが目標となる．

このような問題は，グラフ上での統計的推定問題として考えることができる．グラフ上のノードの性質，例えばタンパク質の機能など，をデータから推定するにはラベル伝播（例えば [1]）という技術が利用でき，これまで一定の成果が報告されている．しかし，多くの場合入力データとして複数のグラフが存在し，それらのうちどれが興味あるタスクに対して重要なのかは事前には明らかでないことが多い．計測技術の発展に伴いこのような問題点が急速に顕在化している．そこで本研究では複数のグラフ上での予測問題を考え，複数の情報源などから得られるグラフを結合し予測精度を向上させる方法論について考察する．

検討内容

複数のグラフを結合する方法としてラベル伝播法の考え方に基づくグラフ結合アルゴリズムを提案した．提案したアルゴリズムの特徴は，ラベル伝播の仮定に基づいた上で，不要と思われるグラフを削除できることにあり，これによって予測精度の向上と解釈性の向上が期待される．ここでの解釈性とは候補グラフ集合のなかから予測にとって重要なものを選びだすことを指しており，結果からデータに関する知見を得るために必要な性質となる．

アルゴリズムの概略を図 1 に示した．アルゴリズムは各グラフに重要度を表す重みを与え，それを反復的に調整する．ここでは数学的な詳細は省略するが，図中の繰り返しを行うことで不要なグラフの重みは最終的に消去される．この性質はスパース性 (sparseness) と呼ばれる統計学上の性質と関連が深いため，提案手法を Sparse Multiple Graph Integration (SMGI) と呼称することとする．

数理的な解析により SMGI の持つ定性的な性質をいくつか明らかにした．前述のスパース性については数値最適化の理論からラベルに対して滑らかなグラフが重みを得て，滑らかでないグラフが削除されることを示した．これは元々のラベル伝播法の考え方と合致する．また，SMGI にはグルーピング効果 (grouping effect) と呼ばれる効果があり，性質の似たグラフは似た重みを与えられることを保証することができる．これは似たグラフは同時に選ばれたり（零でない重みを与えられる），同時に削除されたりしやすいことを示しており（似たグラフは大きく異なる値を取れないため），直観的に非常に自然な結果が得られることを保証する．この性質は有名な統計モデル elastic net[2] の性質と類似しており，数理的にも類似性がある．また，最適化計算は concave-convex

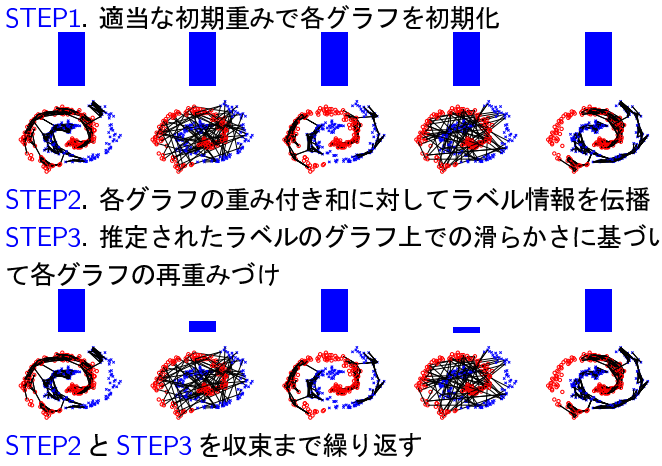

図 1: アルゴリズムの概略．入力となるグラフは各々データに関する異なる情報を持っていたり，非常にノイズが強い場合などもある．提案法は各グラフに対する重みを適応的に調節していく．

procedure(CCCP)[3] という非線形最適化の枠組みに当てはめることができるため効率よく解くことができる．

結果

表 1 はタンパク質の機能予測における AUC による予測精度評価，表 2 はその中で含まれているノイズグラフに対して各手法が零でない重みを与えてしまった割合である．データは [4] から取得したものを利用しており，2 種類の配列情報 (BLAST，Smith-Waterman) と相互作用ネットワークから作成したグラフを利用した．それぞれに対して複数パターンのグラフ生成法を用いたことで結果として今回は計 51 個のグラフが存在する．また，これらのグラフは実際のデータと関係のないノイズグラフを 20 個含んでおり，こういった関係のないグラフを除去できるか評価したのが表 2 である．表中では SMGI 以外の比較手法として OMGSSL[5]，GMKL[6] の結果を掲載した．

表 1: AUC による予測精度評価

SMGI	OMGSSL	GMKL
0.91 (0.04)	**0.91 (0.03)**	0.84 (0.06)

表 2: ノイズグラフを選んだ割合

SMGI	OMGSSL	GMKL
0.07 (0.25)	1.00 (0.00)	0.84 (0.36)

SMGI は高い予測精度と，ノイズ除去率を持っていることが確認できる．AUC については OMGSSL も高い精度をもっている．比較されている手法の中では OMGSSL が SMGI と最も類似しており，この結果はその意味において合理的ではある．ただし，グラフを明示的に選択するには不要なグラフを削除する SMGI が適しており，表 2 の値の差からそれは読みとることができる．

考察

本研究では，生物学データの機械学習アプローチによる解析としてグラフに基づく予測を対象に

し，複数グラフの得られる状況に対して必要なグラフのみを選び出し精度の良い予測を行うための方法を提案した．タンパク質機能予測問題による計算機実験により，提案法が高い精度とノイズ除去力を持つことを実証した（ここまでの成果は [7] において発表を行っている）．一方で，OMGSSL との精度の差や，実験結果の解釈については未だ解析が十分ではなく，より大規模な実験や他のタスクに対する適用も今後の課題としてあげられる．

参考論文

[1] X. Zhu, Z. Ghahramani, and J. Lafferty, “Semi-supervised learning: From gaussian fields to gaussian processes,”in Proceedings of the 20th Annual International Conference on Machine Learning, T. Fawcett and N. Mishra, Eds. AAAI Press, 2003.

[2] H. Zou and T. Hastie, “Regularization and variable selection via the elastic net,” Journal Of The Royal Statistical Society Series B, vol. 67, no. 2, pp. 301.320, 2005.

[3] A. L. Yuille and A. Rangarajan, “The concave-convex procedure,”Neural Computation, vol. 15, pp. 915.936, 2003

[4] G. R. G. Lanckriet, T. D. Bie, N. Cristianini, M. I. Jordan, and W. S. Noble, “A statistical framework for genomic data fusion,” Bioinformatics, vol. 20, no. 16, pp. 2626-2635, 2004.

[5] M. Wang, X.-S. Hua, R. Hong, J. Tang, G.-J. Qi, and Y. Song, “Unified video annotation via multigraph learning,”IEEE Transactions on Circuits and Systems for Video Technology, vol. 19, pp. 733-746, 2009.

[6] A. Argyriou, M. Herbster, and M. Pontil,“Combining graph laplacians for semi.supervised learning, ” in Advances in Neural Information Processing Systems 18, Y. Weiss, B. Schölkopf, and J. Platt, Eds. Cambridge, MA: MIT Press, 2006, pp. 67-74.

[7] 烏山昌幸, 馬見塚拓, ラベル伝播アルゴリズムにおける複数グラフのスパース結合法, 情報論的学習理論と機械学習研究会 (IBISML), 信学技報, vol. 112, no. 279, IBISML2012-58, pp. 171-178, 2012 年 11 月.