The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

アイテムのクラス情報を利用した非負値行列分解 NMF

Non-negative matrix factorization with information expansion using item class information

村上勝彦
Katsuhiko Murakami

東京大学 医科学研究所
The Institute of Medical Science, The University of Tokyo

Non-negative matrix factorization (NMF) is widely used for various problems, such as item recommendation of shopping, image recognition and bioinformatics. However, when the data are sparse, and the row and columns of the items have no association with others, these items tend to be independent resulting poor linking. Here we investigated a method to compensate such too sparse data, by adding information of hierarchical relationships among the items to the matrix analyzed. We show how the additional information helps to make desired new clusters. In addition, too much expansion of class information makes many items into the same clusters, the situation in which is not desired. Some ideas to avoid it are discussed.

1. はじめに

非負値行列分解(Non-negative matrix factorization; NMF)は、顔画像の要素の抽出や単語の分類といったテキストマイニングの方法として提案された[Lee 99]。その後、買い物での商品推薦、画像処理、バイオインフォマティクスなど様々な分野で応用されている。NMF の解析では、1つの非負値の要素からなる行列があたえられ、それを新たな2つの非負値行列の積で近似できるような行列を求める。それによって、もとの行列のなかの、複数行と複数列の中に存在する関連を見出す解析である。

買い物での商品推薦の場合、購入者と商品（アイテム）がそれぞれ行と列になり、購入点数が要素になるデータが解析対象となる。関連付けの根拠は共起頻度にもとづくため、購入事例の少ない個人や商品については、他と関連を見いだせることはあまりない。ほとんど買われない商品が全体の取引の多くを占めるというロングテールという現象でも知られるように、これらの割合は非常に高い。全商品のなかで評価値の予測が可能な商品の割合を被覆率 (coverage) と呼ぶが、ロングテールの状況では被覆率が下がってしまい、多くの商品について推薦ができないという問題があった。

NMF はバイオインフォマティクスの分野でも、遺伝子発現データの解析にも利用されている。NMF は遺伝子発現解析に使われた[Kim 03, Kim 03]。遺伝子とサンプル（細胞や個人を示す）の2次元軸に想定し、各行に遺伝子、各列にサンプルを配置し、行列要素の値が遺伝子発現量である「遺伝子発現行列」がまず作成される。これから行列分解によって、遺伝子セットとその基本発現量を決め、各サンプルはその組み合わせによって説明されるという解釈がされる。これはがん細胞の分類に有効であった[Brunet 04]。ここでは、概要をつかむという目的では有効であったが、他との関連が薄い遺伝子やサンプルを軽視しており、細部の情報を落としているという問題があった。

「遺伝子発現行列」を起点とした場合、この中から協調して発現が近い遺伝子群とサンプル群を取り出すという「バイクラスタリング」[Pontes 15]という手法も発展してきた。この方法に、分解する前後の行列要素に非負という制約を課したものが NMF に対応する。

我々は過去に遺伝子データベースの機能情報同士の関連性を見出すためにNMFを利用した関連抽出を行った [村上 15]。多くの遺伝子と機能の関連が見出されたものの、50％近くを占める多くの用語についてはつながらないままであった。このように NMF ではスパースなデータの場合に、多くの情報が活用できないことが課題である。

本稿では、このような見落としがちな事例の少ないアイテム（用語）についても、他のアイテムの関連を見出すため、アイテムのクラス情報を利用する方法について議論する。各アイテムはなんらかのクラスに属するので、アイテムのオントロジーデータを利用して上位クラスの用語へ一般化していけば、同じクラスとの関連が見いだせると考えられる。以下では、遺伝子と機能の関連性を見出す事例について検討した。

2. データと方法

遺伝子と機能のデータとして、ヒト遺伝子統合データベース H-InvDB (http://h-invitational.jp/) [Takeda 13]を利用した。遺伝子データベースにある記述のうち、遺伝子の機能については統制された用語で基本情報が記述されている。典型的なものとして遺伝子オントロジー（Gene Ontology; GO, GO タームとも呼ばれる）[Blake 2013] がある。ここでは全遺伝子のうち、GO タームが付与されたタンパク質をコードしている遺伝子(12,261遺伝子)を抜き出した。

2.1 遺伝子オントロジー(GO)

GO は用語の集合であり、各用語に少なくとも親子関係（IS-A 関係）が付与されているので、全用語は Directly Acyclic Graph (DAG)構造、階層構造を形成する（図1）。古典的には1遺伝子には1つの機能は1つではなく、複雑な機能を説明するべく、さまざまな観点から関連用語が付与されている。1つの遺伝子には、主要な情報として該当する GO が複数付与されている。あまり調べられていない遺伝子もあり、その場合は何の GO も付与されていない。付与される GO の階層レベルは特に決まりがなく上から下まで使われている。

木構造のルートは 3 種の概念、すなわち機能(molecular function)、生物学的な反応過程（biological process）、および細胞構成要素（cellular component）となっている。本来、生命現象は多くの要素が互いに絡み合っているので、これらの用語や概念は互いに繋がっているはずである。しかし、そのような繋がりはデータベースに陽にはかかれていない。そこで、そのような関係を NMF により発見して補完していく必要がある。

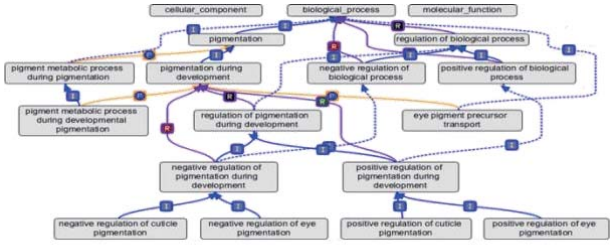

図1　Gene Ontology (GO) の階層構造。GO タームがノード、関係がリンクであり、全体がネットワークで表現されている。

2.2 遺伝子と機能の行列データと行列分解

ヒト遺伝子データベースから取得した、GO タームが付与された 12,261 遺伝子について、縦軸に遺伝子、横軸に用語とする行列を作成し、情報付与の有無を 1,0 で表現する（図2）。

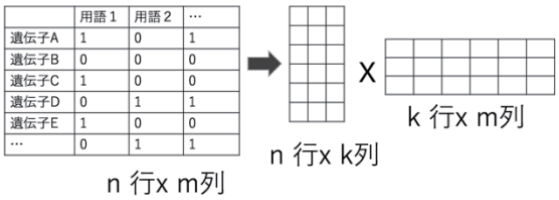

図2　遺伝子と GO タームのデータ構造と行列分解

ユニークな GO ターム回数（異なり数）は 1,741 個であり、延べ数は 40,871 個であった。これに付与されていた 533 個の GO タームを使用した。GO が付与されていた個数は1遺伝子あたり平均 3.3 個であった。通常の文書解析の場合と異なり、1遺伝子にある GO が付与されるのは0回か1回である。

この行列 V を式(1)のように2つの行列の積で近似する。

$$\mathrm{V_{i\mu}} \approx (WH)_{i\mu} = \sum_{a=1}^{k} W_{ia} H_{a\mu} \qquad (1)$$

ここで、V, W, H についてすべての行列要素が非負値である。行列の次元は、V が n×m、W が n×k、H が k×m である。また、k は基底ベクトルの数で、解析の際に固定値として与える。W と H の推定方法としては、以下のようなコスト関数を定義してその最小化を行う。

$$f = \|V - WH\|_F^2 \qquad (2)$$

右辺添字の F はフロベニウスのノルム(V と WH の同じ行と同じ列の要素毎に差をとり、その平方和)である。計算には R ライブラリ Non-Negative Linear Models (NNLM)[Lin 16] を使用した。

3. 関連を見出すための情報の拡張

上記の行列データの間に隠れた関連性を見出すことがここでの目標である。しかし、実際には行列はスパースである。1遺伝子あたりの平均 GO タームの数は3個である[村上 15]。通常の方法では、共起が少ない項目はノイズとして扱われて消えてしまうことになる。

ここでは、各列で要素値の合計（頻度）が t 回以下の低い頻度の GO タームについて、オントロジーの階層情報を参照して階層上位の GO タームを同定する。その GO タームはもともとその遺伝子には付与されていない（下位語が付与されているので上位語は付与する必要がない）。そこで階層が1つ上の上位語について図2の行列のフラグをたてる。こうして背景知識となるオントロジーから関係情報を取得して、データ行列に追加することで上位の GO タームにおける共起頻度が増加し、関連を拾いやすいデータ構造となった。

アイテム推薦の問題においても、商品のカテゴリについての階層データを作成し、同様にデータ拡張が適用できる。

付与した上位の GO タームについても、さらに階層レベルを上げて追加することも考えられるが、階層をあげすぎると、すべての用語について最上位のターム(「molecular function」など)がどの遺伝子にも付与されてしまって意味がない。適切なレベルの階層まで参照することが望ましい。この対策には相互情報量、または最大アイテム数か最大遺伝子数などの基準を設けることが考えられる。

4. おわりに

本稿では、スパースな行列データを NMF で解析する場合に、共起頻度の少ない孤立しがちなアイテム（ここでは GO ターム）について、クラス情報を付与することで、行列データの情報を増やす方法を提案した。これにより、多くの孤立アイテムについて従来は困難であった互いの関連性把握ができるようになった。

今後の課題は、行列要素値から関連遺伝子（および GO ターム）選定の自動化である。係数の大小によって選定するが、閾値を決定する客観的な方法が必要である。将来的には得られた知識データによる推論システムを構築したい。

参考文献

[Blake 13] Blake, J. a, Dolan, M., Drabkin, H., Hill, D. P., Li, N., Sitnikov, D., … Westerfield, M. Gene Ontology annotations and resources. *Nucleic Acids Research*, *41*(Database issue), D530-5. 2013.

[Brunet 04] Brunet, J.-P., Tamayo, P., Golub, T. R., & Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(12), 4164–9. 2004.

[Devarajan 08] Nonnegative matrix factorization: An analytical and interpretive tool in computational biology. *PLoS Computational Biology*, *4*(7). 2008.

[Kim 03] Kim, P. M., & Tidor, B. (2003). Subsystem Identification Through Dimensionality Reduction of Large-Scale Gene Expression Data. *Genome Research*, *13*(7), 1706–1718, 2003.

[Lee 99] Lee, D. D., & Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature*, *401*(6755), 788–791.1999.

[Lin 16] Lin, E. X. NNLM: A package For Fast And Versatile Nonnegative Matrix Factorization. (Web) 2016.

[Pontes 15] Pontes, B., Giráldez, R., & Aguilar-Ruiz, J. S. (2015). Biclustering on expression data: A review. *Journal of Biomedical Informatics*, *57*, 163–180.

[Takeda 13] Takeda, J.-I., Yamasaki, C., Murakami, K., Nagai, Y., Sera, M., Hara, Y., … Imanishi, T. H-InvDB in 2013: an omics study platform for human functional gene and transcript discovery. *Nucleic Acids Research*, *41*(Database issue), D915-9. 2013.

[村上 15] 村上勝彦. 行列因子分解による遺伝子データからの潜在的因子の抽出. In *The 29th Annual Conference of the Japanese Society for Artificial Intelligence*, (pp. 40–42), 2015.