BONSAIによる生物学文献データベースからの知識発見

5K－7

臼坂 慎一[1]　Sim Kim Lan[2]　田中 都子[3]　松野 浩嗣[1]　宮野 悟[2]
[1]山口大学理学部　[2]東京大学ヒトゲノム解析センター
[3]宇部工業高等専門学校

1. はじめに

現在、パン酵母の遺伝子制御ネットワークを同定しようとする研究が進められている[1]。パン酵母はすべてのゲノムが読まれており、遺伝子の個数も6500個と推定されている。遺伝子制御ネットワークとは、遺伝子を翻訳してできるタンパク質の活性化/非活性化という関係を何らかのグラフ表現を使って表そうとするものである。生物学の分野でよく使われる文献データベースのひとつにMEDLINEがあるが、これをひとつひとつ読んで遺伝子関係の書かれてある箇所を見つけていけば、遺伝子制御ネットワークを作ることができる。しかし、Yeastに限っても35000件ものアブストラクトがあり、これは大変な作業である。

そこで我々は、サンプルとしてアブストラクトのうちの一部を取り出し、これらを「遺伝子関係を含むアブストラクト(正の例)」と「遺伝子関係を含まないアブストラクト(負の例)」の2つに分け、これらの正の例と負の例を分ける規則をBONSAI[2]を使って発見することを考えた。この規則をオリジナルの集合に適用して正の例として分類された結果から「真の正の例」を見つける確率は、オリジナルのものを調べて「真の正の例」を見つけるよりも高くなることが期待される。

本稿ではMEDLINEにあるYeast関係のアブストラクトのうち、ジャーナルCellに掲載されている764件について実験を行った結果について述べる。

2. BONSAIによる文献選別の効率化

BONSAIは、文字列集合を対象として、正の例と負の例からそれらを区別する規則を文字の分類（インデキシング）と決定木の形で表現した仮説を見つけ出す機械発見システムである。

我々はこのBONSAIをアブストラクトの解析に応用することを考えた。アブストラクトにおいて遺伝子関係が書かれているセンテンスを正の例、そうでないセンテンスを負の例としてBONSAIに入力として与え、これらを区別する規則を見つける。

BONSAIは文字列を扱う機械学習システムであるが、ここでBONSAIに与える文献は文字の列ではなく単語の列である。ところで文献に使われる単語の中には、正の例のセンテンスによく現れる単語や負の例のセンテンスによく現れる単語が存在すると考えられる。

そこで、アブストラクトの単語1つに対して、文字1つを対応させることを考える。正の例と負の例を区別するのに特徴的に使われそうな単語とそうでなさそうな単語に対して別々の文字を割り当てる。この文字列の集合をBONSAIに入力し、正の例と負の例に分ける規則を見つける。

次に、BONSAIにより見つけ出された規則から文献をどれくらい効率よく選別できるかについて考える。

調べたい集合から適切な方法で一部を取り出してそれを訓練集合と呼び、残りを検査集合と呼ぶ。あらかじめ訓練集合の要素を一つ一つ生物の専門家が正の例(pos)と負の例(neg)に分ける。この訓練集合に対してBONSAIを用いて決定木とインデキシングを得る。BONSAIがposから正の例を取り出す確率をT_p、同様に、negから負の例を取り出す確率をT_nとする。

検査集合には正の例がp個、負の例がn個あったとする。BONSAIは検査集合を

正の例：$pT_p+n(1-T_n)$　個
負の例：$p(1-T_p)+nT_n$　個

と分類する。

Knowledge Discovery in Biology Literature Database by BONSAI, Shin-ichi Usuzaka[1], Sim Kim Lan[2], Miyako Tanaka[3], Hiroshi Matsuno[1], Satoru Miyano[2], [1]Faculty of Science, Yamaguchi University, [2]Human Genome Center, The University of Tokyo, [3]Ube National College of Technology

BONSAI によって正の例として分類されたものから「真の正の例」を見つける確率がもとの検査集合から正の例を見つける確率よりも a 倍よいためには、

$$a = \frac{(p+n)T_r}{pT_r + n(1-T_n)} \quad (1)$$

を満たすような T_p,T_n をもつ決定木とインデキシングが得られればよい。このようなとき BONSAI は人間の作業の手間を軽減するための知識を発見したといえる。

3. 実験

ジャーナル Cell には Yeast 関係の論文のアブストラクトが 764 件掲載されている。我々はまずこれを注意深く読み、211 件の正の例と 533 件の負の例に分けた。さらにこれらを、次のように訓練集合と検査集合に分けた。

	正の例	負の例
訓練集合	100 件	100 件
検査集合	111 件	453 件

次に、訓練集合の正の例と負の例を合わせた 200 件のアブストラクトに出てくる単語を全て列挙すると 3716 語あった。これらのうち、正の例を表すのに特徴的に用いられている単語を定めるために、それらの出現頻度を調べた。

出現個数の多いものでかつその割合が正の例に偏っているものが正の例を特徴づける単語であると考えて、118 個の単語を取り出した。なお、遺伝子名の同定にはスタンフォード大学の Saccharomyces Genome Database Project[3]によって作られた表を用いた。これら 118 語から遺伝子名の 16 個を除いた 102 個をさらに、

x:遺伝子関係を強く表しているもの（8 語）
y:遺伝子関係を弱く表しているもの（39 語）
z:遺伝子関係を表していないもの(55 語)

の 3 つのカテゴリーに分類した。これら以外の語のカテゴリーは「o」とし、さらに遺伝子には全て文字「A」を割り当てて、764 件のアブストラクトの単語ををこれら 5 つの文字で置き換えた(表 1 参照)。このようにして文字変換された正の例と負の例を BONSAI に入力した。

実際の場面では訓練集合に対して格段に多い検査集合を用いることが考えられるので、特に正の例については検査集合をより多くとりたいところであるが、訓練集合をあまり少なくするとよい決定木が得られないので、正の例と負の例ともに 100 件とした。なお、検査集合の数が多くなるにつれて選別の精度が上がるという報告が[4]でなされている。

表 1. 置き換え例

アブストラクト	We show that the yeast HAP1 activator locus encodes a protein that binds in vitro to the upstream activation site, UAS1, of the CYC1 gene (iso-1-cytochrome c). Binding of wild-type HAP1 and truncated HAP1 derivatives to UAS1 is evident in crudely fractionated yeast extracts using the gel electrophoresis DNA binding assay. The binding of HAP1 in vitro, like the activity of UAS1 in vivo, is stimulated by heme. HAP1 binds to region B, one of two portions of UAS1 shown to be important by genetic analysis of the site. Surprisingly, HAP1 binds to the same sequence as a second factor, RC2. Both HAP1 and RC2 bind to the same side of the helix, and make similar but not identical major and minor groove contacts that span two full turns. An additional factor that binds to the second important part of UAS1, the region A factor (RAF), is also identified. A model depicting the interplay of HAP1, RC2, and RAF in the control of UAS1 is presented.
5 つの文字で置き換えたアブストラクト	oooooAoxooooooooooooxooAooooooAooAooxpooooooooooooooooAoooooooooooo oAooooooooxoooooooooooAooooooooooAooooooooooooooooooooooooooooo ooooooxoooooooooooooAoooooooAoo

今回行った実験では訓練集合を BONSAI に入力して得られた決定木とインデキシングに検査集合を入力して、どの程度の選別効率が得られるかを観察した。

訓練集合に対しては 30 回の実験を平均して正の例は約 81.7%, 負の例も 82.5%という確率で選別する決定木を得た。また、検査集合については、正の例 63.3%, 負の例 83.7%であった。訓練集合と比較すると精度は落ちているが、検査集合から正の例を見つける確率よりも約 1.7 倍(式(1)より)の確率で正の例を見つけることができた。

4. おわりに

本稿では、アブストラクト全てを読んで正の例を見つけるよりも、BONSAI によって導き出された規則に基づいて正の例を見つける方が効率よく文献を選別できることを示した。今後は BONSAI による正の例を取り出す精度を上げていく必要がある。

参考文献

[1] Miyano, S., "Knowledge discovery for genome information processing," Technical report of Japanese Society of Artificial Intelligence, SIG-FAI-9701-11 (6/6), 63-68, 1997.

[2] Shoudai, T. et al., "BONSAI Garden:parallel knowledge discovery system for amino acid sequences," *Proceedings of the Third International Conference on Intelligent Systems for Molecular Biology,* AAAI Press, pp.359-366, 1995.

[3] ftp://ftp.genome.ad.jp/db/hgc

[4] Arikawa, S. et al., "A machine discovery from amino acid sequences by decision trees over regular pattern," *New Generation Computing* 11, pp.361-375, 1993.