研究報告

ノンパラメトリックベイズモデルによる無限要素を持つ情報処理システム研究

計算機センター 支援組織　助教　鏑　木　崇　史

本研究の目的は、無限要素を持つノンパラメトリックベイズモデルを用いてデータのクラスタリングや予測精度を高める技術を模索する事である。ノンパラメトリックベイズモデルは機械学習分野において構造未知のデータを適切にクラスタリングできるとして注目を集めている。本研究では（1）文書情報（2）生命情報　の2つについて、無限要素を持つ混合モデルを適用したアルゴリズムを提案することを目標としている。

（1）文書情報

文書情報の具体的なデータとしては学習院コンピュータ支援組織で取り扱った事例データベースを想定している。当該データベースには機材の障害、ソフトウェアの使用方法などの知識が蓄積されている。本研究ではこれらの文書情報を利用して適切な分類の構築を目指している。

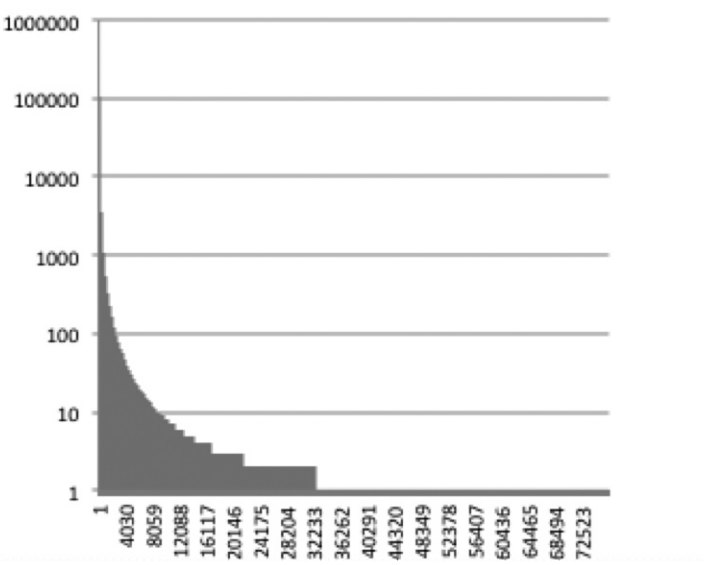

図 1 単語の出現頻度分布

まず、蓄積文書を形態素解析により 1 単語モデルで扱えるように変換することを行った（図 1）。名詞に限定してその頻度分布を解析したところ、ごく少数の名詞が非常に多数出現する Zipf 則がみられた。

さらに、同義語の表現の揺れなどが多数見受けられた。このことから、扱う単語を絞るなど、何らかの事前処理が必要になることが明らかになった。

そこで、比較的重要と思われる単語を抽出し、システム更新時（2012 年 1 月～ 3 月）と平常時（2011 年 10 月～ 12 月）の単語出現頻度の割合を比較した。その結果、これらの単語分布に広がりがあり、文章のクラスタリングに適する単語セットが抽出できたものと思われる。本プロジェクトでは、今後の文書クラスタリングに向けて用いることのできる単語セットの基礎を構築できた。

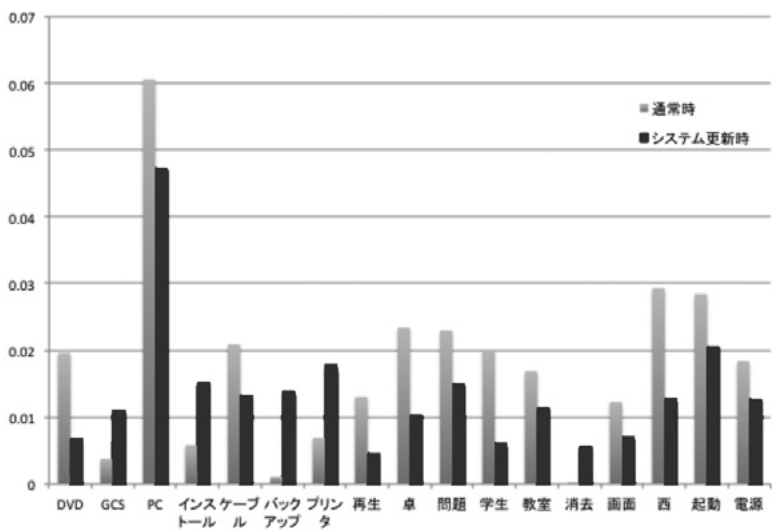

図２ システム更新時と通常時の単語出現頻度の違い

（2）生命情報

具体的なデータとしてはタンパク質の機能データベースを想定した。今日までにさまざまな生物のゲノムが解読され、数多くのアミノ酸配列の特定が進んでいる。一方で、機能が判明しているアミノ酸の数は実験手法の改善などにより進みつつあるが、まだその差は大きく開いている。

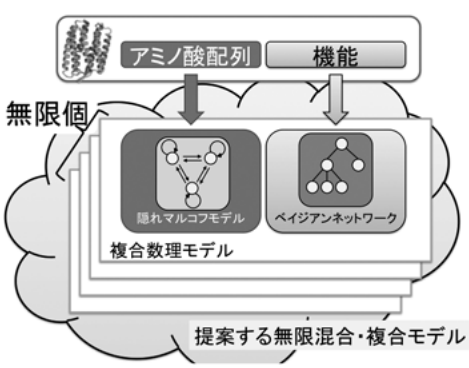

図３　提案する無限混合モデルによるタンパク質機能予測の概念図

この研究ではタンパク質の機能を示す用語集の１つである遺伝子オントロジー（GO）とアミノ酸配列を学習し、機能未知のアミノ酸配列から GO を推定することを目標とした。GO はタンパク質の機能を網羅的に集めた概念の集合体であり、木構造を持っている。機能の概念は木構造の根では最も一般的なものとなっており、葉に近くなればなるほどより具体的な機能を示す用語となっている。このような構造をモデル化するために､GO はベイジアンネットワークモデルを用いた。一方、アミノ酸配列はタンパク質を構成するアミノ酸を空間に関する一次の配列であると考え、隠れマルコフモデル（HMM）でモデル化する。

本研究では、数理モデルのパラメタと、観測されたデータ、ここではアミノ酸配列 y と機能情報 x が与えられたときに、どの程度適合しているかの指標（尤度）が必要になる。HMM のパラメタをΦ、ベイジアンネットワークのパラメタをΨとしたとき、尤度は次のように定義される：

$$P(y|\Phi) = \sum_{z} P(y, z|\Phi) = \sum_{z} P(y|z, \Phi)P(z|\Phi)$$

$$P(x|\Psi) = \prod_{i=1}^{N} P(x_i|\mathrm{Pa}(x_i), \Psi_i)$$

ここで z は隠れ状態列、N は機能の総数、Pa () は変数 χ_i に対する親ノードのすべてである。

本研究では、この複合数理モデルに重み p_k をつけた線形和で示される混合モデルとして扱う。つまり

$$P(x, y|\Phi, \Psi) = \sum_{k=1}^{K} p_k P(x|\Psi_k)P(y|\Phi_k)$$

$$\sum_{k=1}^{K} p_k = 1$$

さらに、タンパク質データはいくつの集合でモデル化するのが適切か不明なため、$K \to \infty$ とするのが適切である。本研究では、Dirichlet Process 事前分布による無限混合モデルを採用した。Dirichlet Process 事前分布には幾つかの表現方法があるが、本研究では Stick-Breaking 表現を用いた。

パラメタΦ、Ψ、p_k の学習には最大事後確率（MAP）期待値最大化法（EM）を用いた。

提案したアルゴリズムの予測精度を比較するため、実験を行った。データセットはタンパク質配列数 378、機能数 153 であり、10 分割交差検定を行った。

ここで、比較手法として Baseline を定義する。この方法では、学習データにおける機能の出現頻度のみを用いたものであり、国際的なタンパク質機能予測コンテストでも比較用に持ちられた方法である。性能の指標としては ROC 曲線 (Receiver Operating Characteristic) を用いて比較する。この方法では、縦軸に True Positive Rate（真陽性）と横軸に False Positive Rate（偽陽性）として

プロットしたものである。ROC 曲線下の面積（Area under the curve，AUC）はアルゴリズムの性能の良さを表す。0 から 1 までの値をとり，完全な分類が可能なときの面積は 1 で，ランダムな分類の場合は 0.5 になる。

性能比較の ROC 曲線を図 4 に示す。提案手法の AUC は 0.91、一方 baseline 手法は 0.79 であった、このことから、提案手法が有用であることが示すことができた。

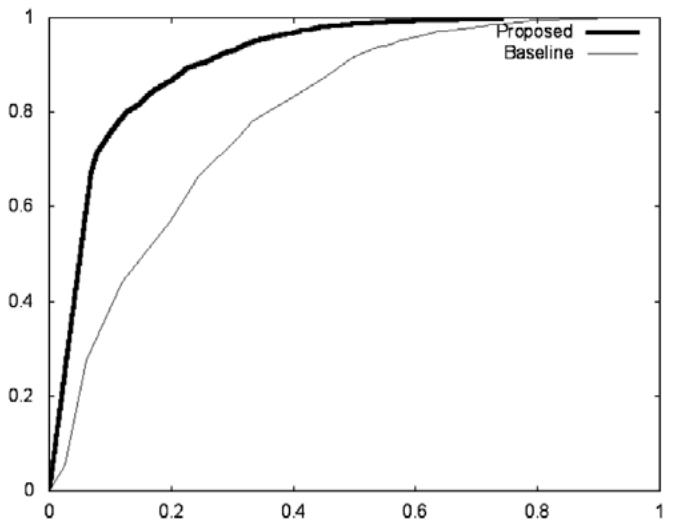

図 4　提案手法（Proposed）と比較手法（baseline）の性能評価

以上のように、本研究では無限要素を用いた情報処理システムの基礎を構築することができた。今後の展望としては、次のようなことが挙げられる：

・文章情報について、本研究ではデータセットを作成する段階で研究期間終了を迎えた。今後は latent Dirichlet allocation などの手法を用いてトピック推定への拡張を行いたい。

・生命情報について、本研究では HMM の隠れ状態数を固定で行ったが、状態数を無限に拡張したモデルも提案されている。今後は無限状態 HMM での実装を目指したい。