7P－7

並列知識獲得システム BONSAI Garden

正代 隆義[†]　宮野 悟[‡]　Michael Lappe[‡]　内田 智之[§]
九州大学理学部　九州大学理学部　九州大学理学部　広島市立大学情報科学部
岡崎 威生[‡]　篠原 歩[‡]
九州大学理学部　九州大学理学部

これまでに，我々は，学習アルゴリズムによる知識獲得システム BONSAI を開発し，主にアミノ酸配列からの知識獲得実験を行なってきた．このシステムは，正の例と負の例からそれらを説明する仮説を学習するもので，これまでの実験で，膜貫通領域とシグナルペプチッド配列に対して，非常に精度の高い小さな仮説を発見し，BONSAI がその能力において大きなポテンシャルをもっていることが判明した．この BONSAI を基本プロセスとして，これを複数個走らせる並列知識獲得システム BONSAI Garden のプロトタイプを作成し，これまで BONSAI で実験を行なってきたものと同じデータを用いて，比較・検討し，その有効性を確認した．

1. はじめに

BONSAI は，構造のわからない１次元の文字列データから知識を獲得するために，正規パターン上の決定木とアルファベットのインデックス化という概念を新たに導入して，[1, 3, 4] 等で得た計算論的学習理論の結果に基づいて開発した機械発見システムである．このシステムは，正の例と負の例からなる文字列の集合が与えられると，それらを分類する仮説として，正規パターン上の決定木とアルファベットのインデックス化を探索する．インデックス化とは，例えば，アミノ酸配列の場合，アミノ酸を表す２０種類の文字を入力文字列の正負の情報を欠落させることなく，あらかじめ設定された，より少ない数の文字に変換する対応付けである．また，正規パターンとは，アミノ酸配列や DNA 配列の「モチーフ」を一般化した概念であり，正規パターン上の決定木とは，各ノードにこうした正規パターンを割り当て，判定規則に用いた決定木である．

これまでの実験で，BONSAI は膜貫通領域とシグナルペプチッド配列に対して，非常に精度の高い小さな仮説を発見し，BONSAI がその能力において大きなポテンシャルをもっていることが判明した．

タンパク質や核酸についての配列データは，その機能や性質により分類されてデータベースに整理されている．しかし，そうしたデータにはノイズが含まれていることが多く，また１つのクラスと分類されているデータも，いくつかの未知のクラスの混ぜ合わせとなっている可能性もある．こうしたノイズを含んだデータや混ぜ合わせのデータに BONSAI を適用すると，インデックス化と正規パターン上の決定木として知識を表現しているため，良い知識の表現が得にくくなる．そこで，BONSAI システムを複数のプロセスとして並列に走らせ，相互にコミュニケーションをとりながら，正の例と負の例から，そのデータについての知識を，データの分類とともに獲得する方式を以下に述べるように構築した．そして，この方式のプロトタイプを，BONSAI Garden というシステムとして実現した．

2. BONSAI システムの概略と機能

BONSAI Garden の核プロセスとなる BONSAI は，文字列データからの知識獲得システムである．BONSAI への入力は，正の例の集合 POS と負の例の集合 NEG である．このシステムは，正の例と負の例からなるこれらの記号列の集合が与えられると，それらを分類する仮説として，アルファベットのインデックス化と正規パターン上の決定木を探索する．正規パターンに現れている記号はインデックス化により変換されたものである．

Parallel Machine Discovery System for Sequences – BONSAI Garden

[†]Takayoshi Shoudai, Department of Physics, Kyushu University, Fukuoka 810, Japan, shoudai@rc.kyushu-u.ac.jp

[‡]Satoru Miyano, Michael Lappe, Takeo Okazaki, Ayumi Shinohara, Research Institute of Fundamental Information Science, Kyushu University, Fukuoka 812, Japan, {miyano,lappe,okazaki,ayumi}@rifis.kyushu-u.ac.jp

[§]Tomoyuki Uchida, Department of Information Sciences, Hiroshima City University, Hiroshima 731-31, Japan, uchida@toc.cs.hiroshima-cu.ac.jp

BONSAI に用いられている主なアルゴリズムは二つからなる．一つは決定木生成機で，もう一つは組み合わせ最適化に用いられている局所探索アルゴリズムである．我々は，Quinlan の ID3 のアイディア [2] を用い，与えられた例に無矛盾な小さな仮説を非常によく見つける効率の良いアルゴリズムを開発した．次に現在得られているインデックス化 I のもとで POS と NEG を変換し，その集合に対してこの決定木の精度評価を行う．そして局所探索アルゴリズムによりインデックス化の変更を行う．このプロセスを，学習に必要な正負の例からのサンプルを変更しながら局所解に落ちるまで続け，その時点でインデックス化と決定木とその精度を出力する．この試行を可能な限り行い，より小さい精度の高い決定木とインデックス化を探索している．

3. BONSAI Garden の方式と概要

BONSAI Garden は，複数の BONSAI をプロセスとして走らせることにより，いくつかの未知の概念が混在した例の集合から，精度の高い複数の仮説を学習しようとするものである．この方式は，正の例の集合が数種類のアミノ酸配列の混ぜ合わせとなっているときに極めて有用である．この方式を大きくとらえると次のように書ける．POS, NEG をほぼ同じ大きさの m 個の部分集合にランダムに分割したものを $\mathcal{P}$, $\mathcal{N}$ とする．

begin /* POS and NEG are given as input */
  let $\mathcal{P}$ be a partition of POS;
  let $\mathcal{N}$ be a partition of NEG;
  repeat
    Classify($\mathcal{P}$,$\mathcal{N}$);
    Merge;
  until *Merge* is impossible
end

分割 $\mathcal{P}$, $\mathcal{N}$ が与えられたとき，*Classify*($\mathcal{P}$,$\mathcal{N}$) は BONSAI を並列に走らせ，新たな分割を構成する．$B_1,\ldots,B_m$ を並列に走る BONSAI とする．各 BONSAI の入力は，正負の例の分割の組 $\langle Pos_i, Neg_i\rangle$ である．各ステージは，2 つの操作からなる．第 j ステージは，次の 2 つの操作が行なわれる．

1. $\langle Pos_i, Neg_i\rangle$ を入力とする BONSAI B_i を走らせ，仮説 (決定木，インデックス化) を得る．
2. 各 BONSAI の間で自身の仮説によってはうまく説明できなかった正，負の例を交換する．

Merge は，こうして得られた分割 $\mathcal{P}$, $\mathcal{N}$ をより小さな分割 $\mathcal{P}'$, $\mathcal{N}'$ に変換する．以下で行なった実験では，正負の各例を *Classify* により得られた m 個の仮説のうち最も小さなものにより説明させることによって，*Merge* を実現している．

4. 計算機実験

数台のワークステーションで BONSAI を走らせ，ネットワークを使ったコミュニケーションを行なうことにより，BONSAI Garden を実現した．

POS としては，GenBank データベースに登録されているいくつかのファイルから合わせて 2,890 個のシグナルペプチッドのアミノ酸配列を取りだし，NEG としてはシグナルペプチッドになっていないアミノ酸配列をランダムに同じファイルから 19,795 個取り出したものを使った．

Bonsai B_0
Indexing:
ACDEFGHIKLMNPQRSTVWXY=11222121202211211101 1
Decision Tree: 122(yes[703,163],no[31,5243])
Accuracy: Positive 95.8%, Negative 97.0%
Bonsai B_1
Indexing:
ACDEFGHIKLMNPQRSTVWXY=00120212100100202221 1
Decision Tree: 01(yes[479,97],21(yes[293,89],no[15,4214]))
Accuracy: Positive 98.1%, Negative 95.8%
Bonsai B_2
Indexing:
ACDEFGHIKLMNPQRSTVWXY=00110222022101202021 1
Decision Tree: 21(yes[522,248],01(yes[202,96],no[30,4853]))
Accuracy: Positive 96.0%, Negative 93.4%
Bonsai B_3
Indexing:
ACDEFGHIKLMNPQRSTVWXY=00121111201121202120 1
Decision Tree: 112(yes[615,163],no[31,4792])
Accuracy: Positive 95.2%, Negative 96.7%

参考文献

[1] S. Arikawa, S. Miyano, A. Shinohara, S. Kuhara, Y. Mukouchi and T. Shinohara, "A machine discovery from amino acid sequences by decision trees over regular patterns", *New Generation Computing* **11**, pp. 361–375, 1993.

[2] J.R. Quinlan, "Induction of decision trees", *Machine Learning* **1**, pp. 81–106, 1986.

[3] S. Shimozono, A. Shinohara, T. Shinohara, S. Miyano, S. Kuhara and S. Arikawa, "Finding alphabet indexing for decision trees over regular patterns: an approach to bioinformatical knowledge acquisition", *Proc. Twenty-Sixth Annual Hawaii International Conference of System Sciences*, pp. 763–772, 1993.

[4] A. Shinohara, S. Shimozono, T. Uchida, S. Miyano, S. Kuhara and S. Arikawa, "Running learning systems in parallel for machine discovery from sequences", *Proc. Genome Informatics Workshop IV*, pp. 74–83, 1993.