BONSAI Garden: Parallel Machine Discovery System for Sequences

Takayoshi Shoudai[1]
shoudai@rc.kyushu-u.ac.jp

Tomoyuki Uchida[2]
uchida@toc.cs.hiroshima-cu.ac.jp

Satoru Miyano[3]
miyano@rifis.kyushu-u.ac.jp

Ayumi Shinohara[3]
ayumi@rifis.kyushu-u.ac.jp

Shinichi Shimozono[4]
sin@ces.kyutech.ac.jp

Takeo Okazaki[3]
okazaki@rifis.kyushu-u.ac.jp

Takeshi Shinohara[5]
shino@ai.kyutech.ac.jp

Satoru Kuhara[6]
kuhara@grt.kyushu-u.ac.jp

Setsuo Arikawa[3]
arikawa@rifis.kyushu-u.ac.jp

[1] Department of Physics, Kyushu University, Fukuoka 810, Japan
[2] Department of Information Sciences, Hiroshima City University, Hiroshima 731-31, Japan
[3] Research Institute of Fundamental Information Science, Kyushu University, Fukuoka 812, Japan
[4] Department of Control Engineering and Science, Kyushu Institute of Technology, Iizuka 820, Japan
[5] Department of Artificial Intelligence, Kyushu Institute of Technology, Iizuka 820, Japan
[6] Graduate School of Genetic Resources Technology, Kyushu University, Fukuoka 812, Japan

1 Introduction

We have developed a machine learning system BONSAI [1, 3, 4] which receives positive and negative examples as inputs and produces a pair of a decision tree over regular patterns and an alphabet indexing as a hypothesis. This system has succeeded in discovering reasonable knowledge on transmembrane domain sequences and signal peptide sequences by computer experiments [1, 3]. However, if data consist of a hodgepodge of sequences, it seems difficult for a single BONSAI system to find a hypothesis of a reasonably small size with high accuracy. Coping with such situation, we have designed a system called BONSAI Garden, which runs several BONSAI systems in parallel, to partition a hodgepodge of data into a small number of disjoint classes together with hypotheses explaining these classes correctly. This paper presents its implementation and demonstration.

2 Method and Overview of BONSAI Garden

The framework of knowledge discovery by a machine learning system BONSAI consists of the following four stages [1, 3, 4]: (1) view design for the data, (2) hypothesis space design for capturing the concepts, (3) designing a learning algorithm, (4) experiments by the learning algorithm for discovering knowledge.

[1]正代 隆義：九州大学理学部物理学科，〒810 福岡市中央区六本松 4-2-1
[2]内田 智之：広島市立大学情報科学部，〒731-31 広島市安佐南区沼田町大塚 151-5
[3]宮野 悟，篠原 歩，岡崎 威生，有川 節夫：九州大学理学部基礎情報学研究施設，〒812 福岡市東区箱崎 6-10-1
[4]下薗 真一：九州工業大学情報工学部制御システム工学教室，〒820 飯塚市川津 680-4
[5]篠原 武：九州工業大学情報工学部知能情報工学科，〒820 飯塚市川津 680-4
[6]久原 哲：九州大学農学研究科遺伝子資源工学，〒812 福岡市東区箱崎 6-10-1

In BONSAI system, the following two basic views are realized: One is a view on the sequences and the other is a view on the alphabet. The "motifs" in amino acid sequences and nucleic sequences can be regarded as a kind of simple regular expressions. This observation leads us to regular patterns as a fundamental view. The other is an alphabet indexing, which is a mapping from an alphabet to a smaller size alphabet that does not lose any essential information from original data. A hypothesis produced by the BONSAI system is a pair of a decision tree over simple regular patterns and an alphabet indexing that classifies the symbols in sequences to a few categories. The learning algorithms we developed for this system are based on the ID3 algorithm in [2] over regular patterns and an algorithm finding a locally optimal alphabet indexing.

We shall briefly show the idea in BONSAI Garden. Let POS and NEG be sets of positive and negative examples, respectively. When POS consists of examples from several unknown concepts, it is not reasonable to try to find a single decision tree with an alphabet indexing that explains POS and NEG. In such case, it is more reasonable to partition the set POS so that each set is explained accurately by a small decision tree with an alphabet indexing. For this purpose, we run multiple BONSAI systems in parallel as follows: Initially, POS is partitioned at random into $m = 2^r$ subsets $\mathcal{P}= (Pos_1, ..., Pos_m)$ of almost even size. Here $m = 2^r$ is chosen appropriately and $POS= Pos_1 \cup \cdots \cup Pos_m$ and $Pos_i \cap Pos_j = \emptyset$ for $1 \leq i \neq j \leq m$. Each BONSAI system B_i takes Pos_i and NEG as a set of positive and negative examples, respectively and find a single decision tree with an alphabet indexing that explains Pos_i and NEG. We modify the procedure for constructing a decision tree so that the depth of the tree does not exceed a fixed constant d. Then, between arbitrary two BONSAIs, we exchange data which can not be explained by the decision tree each other. The next step transforms the partition $\mathcal{P}$ to a smaller partition $\mathcal{P}'$ by merging sets in $\mathcal{P}$. These steps are iterated until no merge is possible.

The above idea is implemented on a network of workstations. This enables to realize the idea of BONSAI Garden, which requires a large amount of computation, so that it runs on a commonly used workstation system with good response time.

3 Further Plan in Progress

We can construct an automatic predictor generator by using BONSAI Garden as a kernel. We assume, as input data, sequences such that some segments are specified on each sequence. Then, with these sequences, the system shall classify the sequences into several classes, and for each of these classes it will generate a program which can predict the specified segments on a sequence. We are now implementing this predictor generator system on a network of UNIX workstations with a high performance graphic machine as an interface.

References

[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.