Parallelized Knowledge Discovery System: An Enhancement to BONSAI

Hideo Bannai [1]
bannai@ims.u-tokyo.ac.jp

Toshio Masuda [2]
tmasuda@ims.u-tokyo.ac.jp

Masao Nagasaki [1]
masao@ims.u-tokyo.ac.jp

Tomohiro Yasuda [1]
tyasuda@ims.u-tokyo.ac.jp

Osamu Maruyama [3]
maruyama@ims.u-tokyo.ac.jp

Satoru Miyano [3]
miyano@ims.u-tokyo.ac.jp

[1] Department of Information Science, Graduate School of Science, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8654, Japan

[2] Course in Computer Science and Information Mathematics, Graduate School of Electro-Communications, 1-5-1 Chofugaoka, Choufu-city, Tokyo 182-8585, Japan

[3] Human Genome Center, Institute of Medical Science, The University of Tokyo, 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan

1 Introduction

BONSAI is a machine learning system for knowledge acquisition from positive and negative examples of strings. It is reported that the system has discovered knowledge which can classify amino acid sequences of transmembrane domains and randomly chosen amino acid sequences located in other parts of the PIR database, with over 90% accuracy [4]. A hypothesis generated by the system is a pair of a classification of symbols called an alphabet indexing, and a decision tree over regular patterns, which classifies given examples with high accuracy. The whole algorithm of the system consists of two parts: a learning algorithm for constructing a decision tree over regular patterns, and a searching algorithm for finding an alphabet indexing to produce a better decision tree.

Through providing a service of BONSAI system, which is available at our web site `http://bonsai.ims.u-tokyo.ac.jp/bonsai/`, we have found problems concerned with the system. One of the problem is that for the size n of an alphabet indexing, which can be specified by the user, the current system is implemented to take time exponential in n. In fact, it is impossible to execute the system with a large n. To overcome this situation, we have discussed how to parallelize BONSAI and succeeded in implementing BONSAI in parallel with some more enhancements.

2 Improvements of BONSAI

Local Search in Parallel:

Generally, an *indexing* ψ *of an alphabet* Σ by another alphabet Γ is a mapping from Σ to Γ. The *neighbors* of ψ are the indexing whose distance from ψ is one, where the distance between ψ and an indexing ϕ of Σ by Γ is defined by $|\{\sigma \in \Sigma \mid \psi(\sigma) \neq \phi(\sigma)\}|$. For an alphabet indexing ψ, the BONSAI system produces a decision tree T by employing a learning algorithm, and outputs as a hypothesis the pair ψ and T. To find the best possible hypothesis, a pair of an alphabet indexing ψ and a decision tree T, the system is designed to find an alphabet indexing ψ by a local search method. In this local search method, for an alphabet indexing ψ, the system finds a local optimum ϕ among the neighbors of ψ, and replaces ψ with ϕ. The search for a local optimum is repeated until ϕ is no better than ψ. We have implemented in parallel this procedure searching for a local optimum by using multi-threads. The improvement in performance concerning with this work is shown in Fig. 1.

Data Filter:

The algorithm of BONSAI has no restriction on strings as input. However, the system has been mainly applied to mere sequence data such as DNA or amino acid sequences. The reason would be that, as positive and negative examples, the system accepts only strings over the ASCII character set, say Σ, and construct an alphabet indexing of Σ, that is, each alphabet of Σ is indexed to a symbol of another alphabet. To conquer the obstacle to handling various data, we have modified BONSAI to execute the preprocess which employs a filter, called a *data filter*, transforming data according to the user's intention. This enables us apply various data to BONSAI just by creating a plug-in of a filter for the data type. Currently, data filters for English and Japanese sentences are available, which implies that BONSAI can try to automatically acquire knowledge from documents.

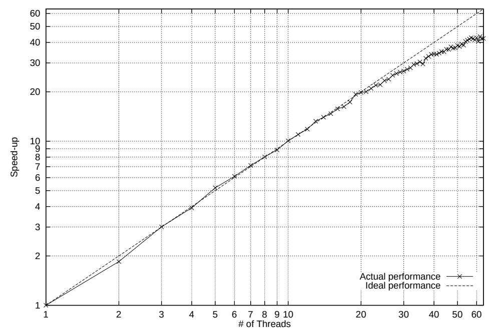

Figure 1: Performance of BONSAI. The performance is measured on a Sun Ultra Enterprise-10000 with 64 processors of 250MHz UltraSPARC. The implemented program adopts the POSIX thread. 689 positive (Transmembrane domain taken from the PIR database) and 19256 negative (generated randomly) examples were used. The window size was set at 10 and the indexing alphabet size was set at 6.

Node Limit:

Some users prefer a compact hypothesis to a precise one. A smaller hypothesis may capture the knowledge involved in a more comprehensible manner. Compact hypotheses in BONSAI are created by limiting the size of the decision tree. Even a *stub*, which is a decision tree with exactly one node, has been verified to be an acceptable model of a hypothesis [1]. Our new BONSAI system provides an option that limits the size of the decision tree created, therefore making a stub available as an alternative hypothesis in the BONSAI system.

Boosting:

Boosting is a general method which can be used to reduce the error of any "weak" learning algorithm, as long as the algorithm can consistently generate classifiers that are better than random guessing [3]. A form of boosting called **AdaBoost** [2] has been empirically shown to improve BONSAI's precision [1]. We have integrated **AdaBoost** into our new system, therefore providing results with higher accuracy. Also, boosting stubs has yielded interesting results [1].

3 Concluding Remarks

As we are now in the process of preliminary experiment, we will report the experimental results at a poster site. The service of the new version of BONSAI is available at our BONSAI site
`http://bonsai.ims.u-tokyo.ac.jp/heibon/`.

References

[1] Bannai H., Boosting BONSAI, *Senior Thesis, Department of Information Science, The University of Tokyo*, 1998.

[2] Freund, Y. and Schapire, R.E., A Decision-Theoretic Generalization of On-line Learning and an Application to Boosting, *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[3] Schapire, R.E., The Strength of Weak Learnability, *Machine Learning*, 5(2):197–227, 1990.

[4] Shimozono, S., Shinohara, A., Shinohara, T., Miyano, S., Kuhara, S., and Arikawa, S., Knowledge Acquisition from Amino Acid Sequences by Machine Learning System BONSAI, *Transactions of Information Processing Society of Japan*, 35(10): 2009–2017, 1994.