Extension and Speed up of Knowledge Discovery System BONSAI

Keisuke Iida (1)
Hideo Bannai (2)
Ayumi Shinohara (1)
Masayuki Takeda (1)
Satoru Miyano (2)

(1) Department of Informatics, Kyushu University 33, Fukuoka 812-8581, Japan
(2) Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1
Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
k-ihda@i.kyushu-u.ac.jp

We extend and speed up the BONSAI system, a machine learning system for knowledge acquisition from positive and negative examples of strings. 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 (strings) to either positive or negative. The algorithm of the system consists of two parts: a learning algorithm for constructing a decision tree over regular patterns, and a local search algorithm for finding a good alphabet indexing for the production of the decision tree. Our focus here is in the improvement of the former, increasing both the speed of hypothesis construction, and the descriptional strength of the generated hypotheses.

Although it has been reported that the system can discover knowledge which can classify certain data sets with fairly high accuracy, in the current implementation, only substring patterns (i.e. whether or not a string pattern appears as a substring of the data string) are searched for, and such patterns may not be powerful enough for distinguishing between positive and negative data of a more complex nature. We present a new version of the BONSAI system which implements several, more powerful variations of patterns, namely, subsequence patterns, episode patterns, and approximate patterns. We also implement an efficient branch-and-bound algorithm for finding the best pattern which distinguishes between the positive and negative data sets.

For each node in the decision tree, the pattern which best distinguishes between the positive and negative examples, in terms of matches, is searched for: i.e. a pattern matches most of the positive examples, but does not match most of the negative examples, or vice versa, is desired. All pattern variations we consider satisfy the condition of, that is, for a pattern of some variation based on the string W, a pattern based on any longer string containing W results in a smaller number of matches against a given set of strings. For such patterns and a conic score function, an upper bound of the score for the longer string may be calculated, and the search can be pruned if the upper bound is less than the current maximum score. For episode patterns, the algorithm of is used to efficiently find the optimal threshold L at the same time. A similar algorithm is also applicable for finding a suboptimal mismatch number K in approximate patterns, and is implemented.

The new BONSAI system has been implemented in the Objective Caml language, a simple but powerful functional language. The source code for BONSAI will be available and distributed at <http://biocaml.org/bonsai/>, under the G NU General Public License.