Alphabet Indexing by Cluster Analysis: A Method for Knowledge Acquisition from Amino Acid Sequences

Hideaki Nakakuni[1]
nakakuni@rifis.kyushu-u.ac.jp

Takeo Okazaki[2]
okazaki@rifis.kyushu-u.ac.jp

Satoru Miyano[2]
miyano@rifis.kyushu-u.ac.jp

[1] Department of Information Systems, Kyushu University
6-1 Kasuga-Koen, Kasuga 816 Japan

[2] Research Institute of Fundamental Information Science, Kyushu University
6-10-1 Hakozaki, Fukuoka 812 Japan

1 Introduction

Knowledge acquisition has been an important topic in Artificial Intelligence and a variety of contributions have been made in various fields where computers can be applied. Genome Informatics is one of the most attracting fields for which knowledge acquisition techniques are strongly expected.

In [3] a knowledge acquisiton system for sequence data has been developed and has shown successful experimental results for amino acid sequences. The input to the system consists of a set P of strings called *positive examples* and a set N of strings called *negative examples* satisfying $P \cap N = \emptyset$. In the case of amino acid sequences, twenty symbols representing amino acid residues constitute the alphabet, say Σ, of input sequences. The task of the system is to find a small hypothesis which explains the positive and negative examples with high accuracy. In [1, 3], two notions are introduced for expressing hypotheses: (1) decision tree over regular patterns, (2) alphabet indexing. A *decision tree over regular patterns* classifies given sequences into two classes *positive* and *negative* by making decisions at nodes. An *alphabet indexing* is a mapping $\psi : \Sigma \to \Gamma$ such that the size of an alphabet Γ is smaller than Σ and the sets $\tilde{\psi}(P)$ and $\tilde{\psi}(N)$ remain disjoint, where $\tilde{\psi}(P)$ (resp. $\tilde{\psi}(N)$) is the set of strings over Γ obtained P (resp. N) by transforming the symbols in Σ by ψ. It is shown in [2] that the problem of finding an alphabet indexing is computationally intractable. By this reason, they introduced *pseudo alphabet indexing* that relaxes the condition so that the size of $\tilde{\psi}(P) \cap \tilde{\psi}(N)$ is sufficiently small.

A local search algorithm is devised for finding a pseudo alphabet indexing since the problem of finding an exact alphabet indexing has been shown NP-complete. However, the local search algorithm in [3] is not fast enough when then number of training examples gets larger. This paper developes a very efficient method for finding an pseudo alphabet indexing by employing a cluster analysis technique called Ward's method. The experiments on transmembrane sequences and signal peptide sequences show that this method is surprizingly efficient and could find good alphabet indexings. The following section provides an overview of the method.

[1] 中國秀章：九州大学総合理工学研究科情報システム学専攻，〒816 春日市春日公園 6-1

[2] 岡崎威生，宮野　悟：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

2 Alphabet Indexing by Ward's Method

Cluster analysis is a statististical method for multivariate model in which aim is to see individuals fall into groups or clusters. The clustering depends on the similarity between individuals or clusters that is defined using plural characteristics. The procedure of clustering has a variety of methods depending on the definition of similarity or distance metric. Since no prior information for our data are available in our case, we apply Ward's method [4, 5] which has been known to give a useful classification empirically. In Ward's method, the distance between two clusters is defined so that the sum of squares from the objects to the joint cluster mean minus the sum of squares from the objects to their individual cluster means.

Each object can be expressed as a d-dimensional real vector $x_i = (x_i^1, x_i^2, \ldots, x_i^d)'$ $(i = 1, 2, \ldots, n)$, where d corresponds to the number of characteristics and n corresponds to the number of objects. A cluster C_k is a subset of R^d, where $\bigcup C_k = \{x_1, x_2, \ldots, x_n\}$, $C_i \cap C_j = \emptyset \quad (i \neq j)$. The cluster mean $\bar{x}^j_{(C_k)} = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i^j$ is given by the average and the loss of information E_{C_k} is defined as follows:

$$E_{C_k} \;=\; \sum_{x_i \in C_k} \sum_{h=1}^{d} (x_i^h - \bar{x}^h_{(C_k)})^2.$$

Then the distance between two clusters is given by

$$\begin{aligned} \Delta E(C_i, C_j) &= E_{C_i \cup C_j} - (E_{C_i} + E_{C_j}) \\ &= \frac{|C_i||C_j|}{|C_i| + |C_j|} \sum_{h=1}^{d} (\bar{x}^h_{(C_i)} - \bar{x}^h_{(C_j)})^2. \end{aligned}$$

With this $\Delta E(C_i, C_j)$, we can constract the distance matrix $S = (s_{ij})$. At the initial stage, each x_i becomes a cluster which consists of a single element. We find the smallest distance between clusters, fuse them, and reconstruct the distance matrix. While $\Delta E(C_i, C_j)$ is larger than a given real number or the number of clusters is 1, these steps are repeated. Finally, by assignning a distinct symbol to each cluster, we can obtain a pseudo alphabet indexing.

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*, Vol. 11 pp. 361–375, 1993.

[2] S. Shimozono and S. Miyano, "Complexity of alphabet indexing", to appear in *IEICE Trans. Information and Systems*, Vol. E77-D, 1994.

[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] D. Wishart, "An algorithm for hierarchical classifications", *Biometrics*, Vol. 25, pp. 165–170, 1969.

[5] G. A. F. Seber, "Multivariate Observations", John Wiley & Sons, 1984.