ICM 1998, Berlin, Aug. 18–27

Abstracts of Short Communications and Posters

Section:

16. Applications

1991 MS Classification:

Qian, Minping, School of Mathematics, Department of Probability and Statistics, Peking University

Hidden Markov Model, Context free Grammar and their Applications

In the recent years, the rapid development of computers made it possible to solve large scale complex practical problems, such as DNA sequencing , pattern recognitions of hand writing Chinese characters, speech recognitions. One of the key-steps is to give a formal, general and flexible modeling scheme, which is mathematically well defined so that it can be easily translated into computer programs and also allows to use intuitive ideas and experience to specify and simplify the problems from general setting.

The formal language is introduced to describe and to analyze complex systems and the context free languages allow to specify the relation between "words" by grammar rules with simple regulations. If probabilities are assigned to the grammar rules, then all grammar rules in fact define a random map which maps the set of words to itself. The property of context free makes independence of these maps for each pair of the nearest "words" in a sentence. Thus, if we take the "time" index going along the sentence word by word, we can also consider the sentence as a sample path of a Markov chain (model). From this point of view, the machinery of Markov processes applies. Experiments by the author's group in Peking University for the hand-writing Chinese character show the great potential of this approach. It is also considered having very good performence of hidden Markov models for DNA sequencing. Let us look at the DNA sequence coding. As it is well known that all informations to build the body of a human being or an animal are contained in the DNA sequences of genomes (chromosomes). A world wide grand project which is considered to be very important both theoretically and practically, is carrying to "read" the human genomes. But what can be done up to the level of present technology is that to break the long sequence of close to 100 million bases into shorter sequences of several hundred bases and read out each base of these shorter sequences. In this process the reading error, i. e. misreading, insert fake bases and base missing, is not avoidable. The goal is to know the whole sequence of genomes, and to understand the function of the protein coding by each part of the genome sequences. The Hidden Markov chain in multi-level is an excellent tool for modelling the complicate structure of DNA sequence, and for using as much extra information as biologists can provide as well.