Sequence Data Mining Techniques and Applications

Sunita Sarawagi

Abstract

Many interesting real-life mining applications rely on modeling data as sequences of discrete multi-attribute records. Mining models for network intrusion detection view data as sequences of TCP/IP packets. Text information extraction systems model the input text as a sequence of words and delimiters. Customer data mining applications profile buying habits of customers as a sequence of items purchased. In computational biology, DNA, RNA and protein data are all best modeled as sequences.

Classifying, clustering and characterizing such sequence data presents interesting issues in feature engineering, discretization and pattern discovery. In this seminar we will review techniques ranging from item set counting, MDL-based discretization and Markov modeling to perform various supervised and unsupervised pattern discovery tasks on sequences. We will present case studies from network intrusion detection and DNA sequence mining to illustrate these techniques.

Sunita Sarawagi researches in the fields of databases, data mining, machine learning and data warehousing. She is a member of the faculty at IIT Bombay. Prior to that she was a research staff member at IBM Almaden Research Center. She got her PhD in databases from the University of California at Berkeley.