Chapter 4
A Review of Kernel Methods Based Approaches to Classification and Clustering of Sequential Patterns: Sequences of Discrete Symbols

T. Veena
Indian Institute of Technology Madras, India

A. D. Dileep
Indian Institute of Technology Madras, India

C. Chandra Sekhar
Indian Institute of Technology Madras, India

ABSTRACT

Pattern analysis tasks on sequences of discrete symbols are important for pattern discovery in bioinformatics, text analysis, speech processing, and handwritten character recognition. Discrete symbols may correspond to amino acids or nucleotides in biological sequence analysis, characters in text analysis, and codebook indices in processing of speech and handwritten character data. The main issues in kernel methods based approaches to pattern analysis tasks on discrete symbol sequences are related to defining a measure of similarity between sequences of discrete symbols, and handling the varying length nature of sequences. We present a review of methods to design dynamic kernels for sequences of discrete symbols. We then present a review of approaches to classification and clustering of sequences of discrete symbols using the dynamic kernel based methods.

DOI: 10.4018/978-1-4666-3604-0.ch004

INTRODUCTION

Kernel methods for pattern classification and clustering were presented in the previous chapter. We also explained the design of dynamic kernels for sequences of continuous feature vectors. In this chapter, we present a review on the design of dynamic kernels for discrete symbol sequences.

A discrete symbol sequence comprises of a sequence of symbols, belonging to an alphabet $\sum$, observed or recorded during a process. For example, in coin tossing experiment, the observations being either head (*H*) or tail (*T*) may result in a discrete observation sequence *HHTHTTH*. Here the alphabet $\sum$ is a set of two symbols {*H, T*} and the length of the observation sequence is seven. The discrete observation sequence obtained in another coin tossing experiment may be *THHHTTHHT* resulting in a sequence of different length. One major issue in handling discrete symbol sequences is that the observation sequences are of varying length in nature. This applies to any sequence data. A major source of discrete symbol sequences is the biological sequences such as protein sequences, DNA sequences and RNA sequences. The DNA sequences are strings over four nucleotides, represented by the alphabet $\sum=\{A,C,G,T\}$. The RNA sequences are strings over the alphabet $\sum=\{A,C,G,U\}$. The symbols in the alphabet for DNA and RNA correspond to the following nucleotides: *A*(adenine), *C*(cytosine), *G*(guanine), *T*(thymine), and *U*(uracil). The positions of occurrence of these nucleotides in the chain molecule of DNA or RNA signify the functioning of that DNA or RNA. An example DNA sequence of length 50 is *ATAATAAAAAATAAAAATA-AAAAAAATTAAAAAATATTAAAAAATAAAAA*. Protein sequences are strings over an alphabet of 20 amino acids which are the building blocks of proteins. The kinds of amino acids occurring, their frequency of occurrence and their relative positions of occurrence in a protein sequence influence the functionality of a protein. An example of a protein sequence is *MGTPTLAQPV-VTGMFLDPCH*. Discrete symbol sequences are also used to analyze text data.

A paragraph is considered as a sequence of words. In text analysis, words are the observation symbols derived from a vocabulary of all the words. Discrete observation sequences are also derived by vector quantization of the continuous feature vector sequences extracted from speech data and online handwritten character data. Pattern analysis tasks involving discrete symbol sequences are classification and clustering. In order to use kernel methods for these tasks, it is necessary to address the issue of handling the varying length nature of sequences. In some approaches, an explicit feature map (Ding & Dubchak, 2001; Jaakkola et al., 2000; Leslie et al., 2002; Leslie & Kuang, 2003; Liao & Noble, 2002; Lodhi et al., 2002; Logan et al., 2001) is used to obtain a fixed length representation for each of the varying length sequences. In some other approaches, a kernel is designed directly from the varying length sequences (Saigo et al., 2004; Tsuda et al., 2002; Vert et al., 2004; Watkins, 1999). Kernels designed using any of these two methods are called as dynamic kernels for discrete symbol sequences.

The organization of the rest of the chapter is as follows: The next section describes the methods for the design of dynamic kernels for discrete symbol sequences. Then a review of kernel methods based approaches to sequential pattern analysis involving discrete symbol sequences is presented.

DESIGN OF DYNAMIC KERNELS FOR DISCRETE SYMBOL SEQUENCES

The main issue in designing a kernel for discrete observation symbol sequence is to handle the varying length nature of the sequences. The varying length sequences of discrete observation symbols may be explicitly mapped onto a fixed dimensional feature vector and then the kernel is computed as an innerproduct in that fixed dimensional space.