Stochastic Sliding Windows For Sequential Data Mining

Jordan Genoff

Windowing is one of the substantial tools in a principle class of sequential data mining approaches. In general, non-overlapping or overlapping sliding windows are used to split a longer stream into a set of shorter sequences, which set is then treated as an input to some database mining method. Thus significant in some sense sequential patterns or rules are discovered. This paper deals with the case of discrete finite alphabet data streams, where windows, though being fixed or varying in size and/or sliding step, are always dense structures of masking flags. Thus all sequences resulting from the windowing are exact dense subsequences of the source data.

There are cases where the sought patterns or rules in the sequential data are partially stochastic in nature or are hidden by some stochastic interference. A windowing technique is proposed that alone itself is able to enhance the performance of mining in such cases by producing more adequate windowed subsequences. Given the window size and sliding step, a unique non-dense masking structure of flags ("window ") with the same boundary size is used at each windowing step. The structure is produced randomly according to a given probability distribution of on/off flag values in the window frame. The windowed subsequence may be treated as a gapped (where flags in the window are off) one or may be condensed by concatenation of the non-gap regions (where flags are on). Such a formulation implies that more than one windowing actions with random windows must be taken at each windowing step in the input data stream. The on/off probability distribution in the window frame and the number of repetitions at each step are main parameters of the setup, along with the window size and sliding step.

The setup is thoroughly investigated as a Monte Carlo one and is experimentally tested on synthetic and real-life data of biological nature.