Synthetic Dataset Generator

We developed a new sequence data generator. It supports a number of input parameters that allow us to control the key properties of the generated sequence datasets as listed in Tab. 1. These include the number of sequences (devices), the average length of each sequence, the number of event types, the number of frequent patterns, the number of outlier patterns, etc. In particular we inject random noises, that is, the events not belonging to any frequent pattern or outlier pattern to mimic the real-world data.

Table 1: Input Parameters to Sequence Data Generator.

Symbol	Description
$\|D\|$	Number of devices (Number of sequences)
$\|S\|$	Average length of sequences
$\|E\|$	Number of event types
$\|F\|$	Number of frequent patterns
$\|O\|$	Number of outliers
$\|L\|$	Average length of frequent patterns
e	noise rate

The generator is composed of two parts, namely *pattern generation* and *sequence generation*.

During the pattern generation phase, pattern candidates, frequent patterns and outlier patterns will be generated. $|P|$ pattern candidates will be generated first. The lengths of the $|P|$ patterns roughly follow Gaussian distribution with $\mu = \frac{2|L|}{3}$ and $\sigma = \frac{|L|}{5}$.

Once the length is determined to be l, the pattern candidate will be generated by randomly selecting l distinct elements from the alphabet A. Then the frequent patterns are generated by randomly select $|F|$ patterns from the pattern candidates. Finally, outliers are obtained by transforming $|O|$ frequent patterns. The transformation includes adding new elements, deleting elements, exchanging positions of elements. A new transformed pattern can be included in the outlier pattern set if it is not identical with any pattern candidates or existing outlier patterns.

Then during the sequence generation phase, the sequence of each device can be generated based on these generated pattern candidates, frequent patterns and outlier patterns. The sequence generation mainly consists of four parts, namely adding outliers, adding frequent patterns, adding unexpected frequent patterns & infrequent patterns, and adding random noise. The process starts with adding $|O|$ outliers to sequences. Based on our definition, outlier patterns are frequent locally in one single sequence, but not global frequent. Therefore, given one outlier, the generator randomly selects one sequence from D, and attaches the outlier pattern for $minLS \leq |LS| \leq 3 \times minLS$ times to the sequence. After adding outliers, frequent patterns are added. A pattern has to be local frequent (appears at least $minLS$ times in one sequence) in at least $minGS$ sequences to be global frequent. Therefore, for each frequent pattern, we randomly select $minGS \leq |GS| \leq 3 \times minGS$ sequences from $|D|$ and add the frequent pattern to each sequence for $minLS \leq |LS| \leq 3 \times minLS$ times. Here, the local support values as well as global support values of outlier and frequent patterns are designed not to be identical but within a given range to better simulate real world datasets. Furthermore, to introduce unexpected frequent patterns and infrequent patterns, we further enrich each sequence with patterns from pattern candidates. To be specific, if the sequence length $|S_i|$ does not reach $0.9 \times |S| \leq |S_i| \leq 1.1 \times |S|$, the generator selects a set of patterns from pattern candidates until the sequence has length $0.9 \times |S| \leq |S_i| \leq 1.1 \times |S|$. The selection also follows Gaussian distribution. That is, the generator gives each pattern in candidate set a numerical label from 1 to $|P|$. The patterns that have labels between $\frac{|P|}{2} \pm \frac{|P|}{5}$ have higher chance to be selected and thus have higher chance to be unexpected frequent patterns. Finally, the generator also introduces random noise to the sequences. Specifically, each element in the sequence has e possibility to be removed.

In the experiments, we generate synthetic datasets with various combinations of sequence length $|S|$, alphabet Size $|A|$, number of frequent patterns $|P|$ and maximum length of patterns $|L|$.