Evaluation of SAX functionalities

Alessio Bertone

Danube University Krems
Department of Information and Knowledge Engineering (ike)

IKE-TR-2007-02 July 2007

Authors: **Alessio Bertone**

Alessio.Bertone@donau-uni.ac.at
http://www.donau-uni.ac.at/ike

Contact: **Danube University Krems**
Department of Information and Knowledge Engineering (ike)
Dr.-Karl-Dorrek-Str. 30
3500 Krems
Austria, Europe
T +43 (2732) 893 - 2453
F +43 (2732) 893 - 4450
ike@donau-uni.ac.at
http://www.donau-uni.ac.at/ike

Table of Contents

1 Introduction

The aim of this report is to illustrate the main features of SAX and to outline the limitations in order to apply this approach to DisCō.

The report is organized as follows: firstly a description of SAX is given (Par. 2), then a discussion illustrated which are the limitations of SAX within the context of DisCō (Par. 3).

2 What is SAX

SAX (Symbolic Approximation Aggregation) [1] [2] is a novel approach to a symbolic representation of time series. It allows lower-bounding distance measures to be defined on the symbolic space, so that it is possible to pass from a common representation of time series (a sequence of data points interpolated by a line) to a symbolic one.

More in details, SAX performs a discretization in two steps (Figure 1). Firstly, a time series is divided into *w* equal-sized segments; the values in each segment are then approximated and replaced by a single coefficient, which is their average. Aggregating these w coefficients form the piecewise aggregate approximation (PAA) representation of the time series. Secondly, in order to convert the PAA coefficients to symbols, the breakpoints that divide the distribution space into *a* equiprobable regions are calculated, where *a* is the alphabet size specified by the user. In other words, the breakpoints are determined such that the probability of a segment falling into any of the regions is approximately the same. If the symbols were not equi-probable, some of the substrings would be more probable than others (as a consequence, a probabilistic bias in the process is added).

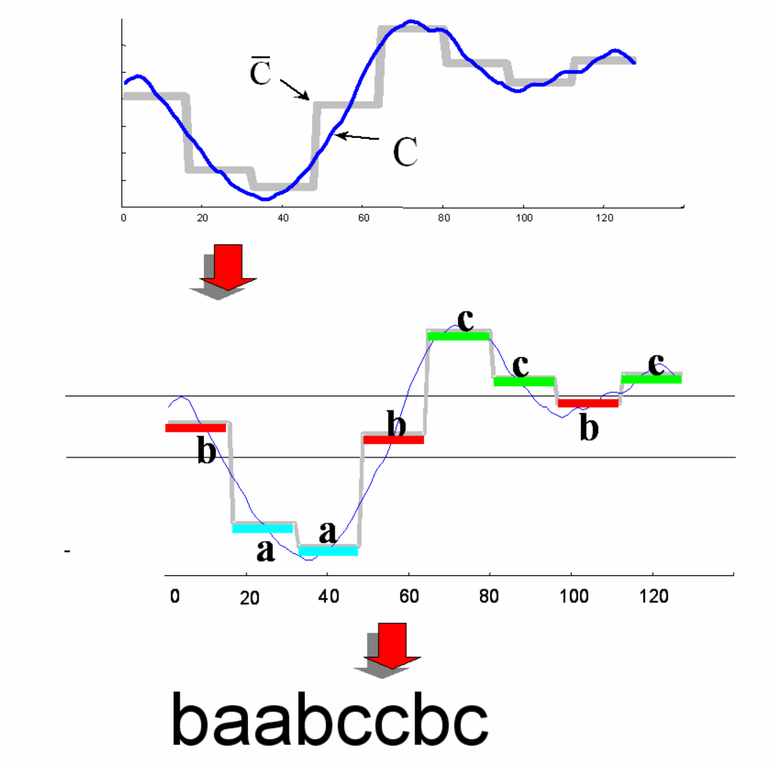

Figure 1: The two steps discretization process of Sax

3 Limitations of SAX in the DisCō context

The following considerations are based on reading and study of Eamonn Keogh´s methods and the attempt to apply them to the context of DisCō.

In order to study the potentialities and the possible limitations of SAX, the proper software has been downloaded (http://www.ise.gmu.edu/~jessica/sax.htm) and installed along with Matlab, since this version has been implemented using this language.

Basically, the aim of the DisCō is to find trends, patterns and relations among patterns within temporal data, which are also multivariate, multidimensional, multigranular and irregularly sampled. In other words, the aim is the discovery of interesting patterns within one or more time series, and the relations among them, also providing the capabilities to deal with different granularities.

Starting from the available software and applying it to the data from DisCō (Figure 2 and Figure 3), here it follows a list of the main limitations encountered:

- Capability to read only data point (no interval, no nominal data)
- Need of a pre-processing (starting from i.e. csv data, a column has been selected, then any possible comma has been deleted)
- Size of the alphabet to be used for the symbolic representation (at the moment 10)
- Representation of the time series on a plot without any time reference, but rather as a set of discretely ordered values
- No possibility to change dynamically the size of the windows / segments, according to which the time series is approximated (this feature seems to be provided in a newer version of VizTree - http://www.ise.gmu.edu/~jessica/viztree/viztree_demo.htm [3], [4])
- Missing data are not treated
- Multiple granularities are not considered
- Two time series can be treated at a time (within VizTree)
- Only explorative analysis (no confirmative, i.e., no possibility to find a given or known a priori pattern)

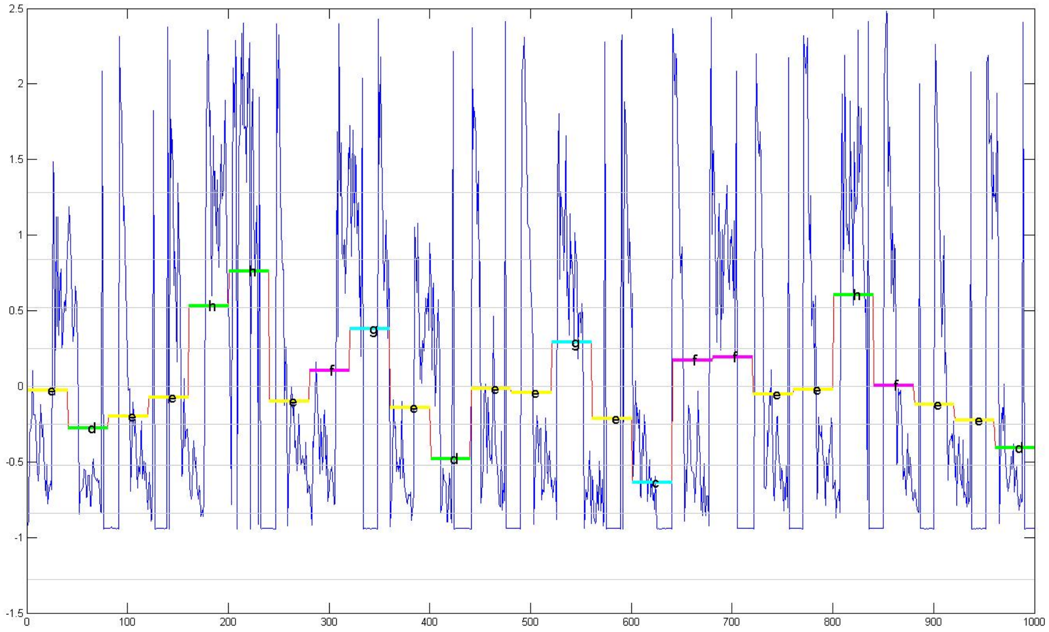

Figure 2: SAX applied to DisCō data (number of segments: 25, alphabet size: 10)

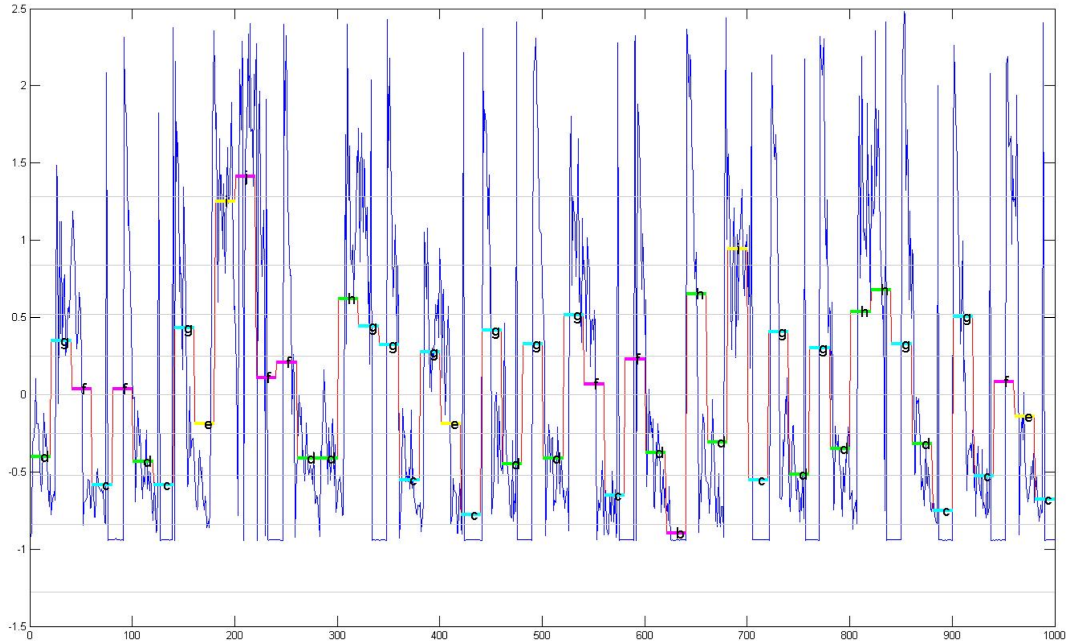

Figure 3: SAX applied to DisCō data (number of segments: 50, alphabet size: 10)

4 References

[1] Lin, J., Keogh, E. & Lonardi, S. (2005). Visualizing and Discovering Non-Trivial Patterns in Large Time Series Databases. Information Visualization Journal.

[2] Lin, J., Keogh, E., Lonardi, S. & Chiu, B. (2003). A Symbolic Representation of Time Series, with Implications for Streaming Algorithms. In proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery. San Diego, CA. June 13.

[3] Lin, J., Keogh, E., Lonardi, S., Lankford, J. P. & Nystrom, D. M. (2004). Visually Mining and Monitoring Massive Time Series. In proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Seattle, WA. Aug 22-25.

[4] Lin, J., Keogh, E., Lonardi, S., Lankford, J. P. & Nystrom, D. M, Viztree: a tool for visually mining and monitoring massive time series databases. In Proceedings of International Conference on Very Large Data Bases, pages 1269--1272, 2004.