Towards Discovering the Intrinsic Cardinality and Dimensionality of Time Series Using MDL

Bing Hu$^1$, Thanawin Rakthanmanon$^1$, Yuan Hao$^1$, Scott Evans$^2$, Stefano Lonardi$^1$, and Eamonn Keogh$^1$

$^1$Department of Computer Science & Engineering, University of California, Riverside, CA 92502, USA
$^2$GE Global Research
{bhu002,rakthant,yhao002}@ucr.edu, evans@ge.com, {stelo,eamonn}@cs.ucr.edu

Abstract. Most algorithms for mining or indexing time series data do not operate directly on the original data, but instead they consider alternative representations that include transforms, quantization, approximation, and multi-resolution abstractions. Choosing the best representation and abstraction level for a given task/dataset is arguably the most critical step in time series data mining. In this paper, we investigate techniques that discover the natural intrinsic representation model, dimensionality and alphabet cardinality of a time series. The ability to discover these intrinsic features has implications beyond selecting the best parameters for particular algorithms, as characterizing data in such a manner is useful in its own right and an important sub-routine in algorithms for classification, clustering and outlier discovery. We will frame the discovery of these intrinsic features in the Minimal Description Length (MDL) framework. Extensive empirical tests show that our method is simpler, more general and significantly more accurate than previous methods, and has the important advantage of being essentially parameter-free.

Keywords: Time Series, MDL, Dimensionality Reduction.

1 Introduction

Most algorithms for indexing or mining time series data operate on higher-level representations of the data, which include transforms, quantization, approximations and multi-resolution approaches. For instance, Discrete Fourier Transform (DFT), Discrete Wavelet Transform (DWT), Adaptive Piecewise Constant Approximation (APCA) and Piecewise Linear Approximation (PLA) are models that all have their advocates for various data mining tasks and each has been used extensively [3]. However the question of choosing the best abstraction level and/or representation of the data for a given task/dataset still remains open. In this work, we investigate this problem by discovering the natural intrinsic model, dimensionality and (alphabet) cardinality of a time series. We will frame the discovery of these intrinsic features in the Minimal Description Length (MDL) framework [11] [21] [28]. MDL is the
cornerstone of many bioinformatics algorithms [7][20], but it is arguably underutilized in data mining [9].

The ability to discover the intrinsic dimensionality and cardinality of time series has implications beyond setting the best parameters for data mining algorithms, as characterizing data in such a manner is useful in its own right to understand/describe the data and an important sub-routine in algorithms for classification, clustering and outlier discovery. To illustrate this, consider the three unrelated datasets in Fig. 1.

Fig. 1. Three unrelated industrial time series with low intrinsic cardinality. I) Evaporator (channel one). II) Winding (channel five). III) Dryer (channel one).

The number of unique values in each time series is, from left to right, 14, 500 and 62. However, we might reasonably claim, that the intrinsic alphabet cardinality is instead 2, 2, and 12 respectively. As it happens, an understanding of the processes that produced these data would perhaps support this claim [10]. In these datasets, and indeed in many real-world datasets, there is a big difference between the actual and intrinsic cardinality. Similar remarks apply to dimensionality.

Before we define more precisely what we mean by actual versus intrinsic cardinality, we should elaborate on the motivations behind our considerations. Our objective is generally not simply to save memory: if we are wastefully using eight bytes per time point instead of using the mere three bytes made necessary by the intrinsic cardinality, the memory space saved is significant, but memory is getting cheaper every day, which is rarely a bottleneck in data mining tasks. There are instead many other reasons why we may wish to find the true intrinsic model, cardinality and dimensionality of the data:

- There is an increasing interest in using specialized hardware for data mining [24]. However, the complexity of implementing data mining algorithms in hardware typically grows super linearly with the cardinality of the alphabet. For example, FPGAs usually cannot handle cardinalities greater than 256[24].
- Some data mining algorithms benefit from having the data represented in the lowest meaningful cardinality. As a trivial example, in the stream: .0, 0, 1, 0, 0, 1, 0, 0, 1, we can easily find the rule that a ‘1’ follows two appearances of ‘0’. However, notice that this rule is not apparent in this string: .0, 0, 1.0001, 0.0001, 0, 1, 0.000001, 0, 1 even though it is essentially the same time series.
- Most time series indexing algorithms critically depend on the ability to reduce the dimensionality [3]or the cardinality [16]of the time series (or both[1][2]) and searching over the compacted representation in main memory. However, setting the best level of representation remains a black art.
- In resource-limited devices, it may be helpful to remove the spurious precision induced by a cardinality/dimensionality that is too high (see example below).
- Knowing the intrinsic model, cardinality and dimensionality of a dataset allows us to create very simple outlier detection models. We simply look for data where