We Have Seen the Future, and It Is Symbolic (Abstract of Invited Talk)

Eamonn Keogh **Jessica Lin** **Stefano Lonardi** **Bill Chiu**

University of California - Riverside
Computer Science & Engineering Department
Riverside, CA 92521, USA
{eamonn, Jessica, stelo, bill}@cs.ucr.edu

Many high level representations of time series have been proposed for data mining. One representation that the data mining community has not considered in detail is the discretization of the original data into symbolic strings. At first glance this seems a surprising oversight. There is an enormous wealth of existing algorithms and data structures that allow the efficient manipulations of strings. Such algorithms have received decades of attention in the text retrieval community, and more recent attention from the bioinformatics community. Some simple examples of tools that are not defined for real-valued sequences but are defined for symbolic approaches include hashing, Markov models, suffix trees, decision trees etc.

There is, however, a simple explanation for the data mining communitys lack of interest in string manipulation as a supporting technique for mining time series. If the data are transformed into virtually any of the other popular data mining representations, then it is possible to measure the similarity of two time series in that representation space, such that the distance is guaranteed to lower bound the true distance between the time series in the original space. This simple fact is at the core of almost all algorithms in time series data mining and indexing. However, in spite of the fact that there are dozens of techniques for producing different variants of the symbolic representation, there is no known method for calculating the distance in the symbolic space, while providing the lower bounding guarantee.

In addition to allowing the creation of lower bounding distance measures, there is one other highly desirable property of any time series representation, including a symbolic one. Almost all time series datasets are very high dimensional. This is a challenging fact because all non-trivial data mining and indexing algorithms degrade exponentially with dimensionality. For example, above 16-20 dimensions, index structures degrade to sequential scanning. None of the symbolic representations that we are aware of allow dimensionality reduction. There is some reduction in the storage space required, since fewer bits are required for each value, however the intrinsic dimensionality of the symbolic representation is the same as the original data.

In this work we introduce a new symbolic representation of time series. Our representation is unique in that it allows dimensionality/numerosity reduction, and it also allows distance measures to be defined on the symbolic representation that lower bound corresponding popular distance measures defined on the original data. As we shall demonstrate, the latter feature is particularly exciting because it allows one to run certain data mining algorithms on the efficiently manipulated symbolic representation. We will demonstrate the utility of our representation on classification, clustering, indexing, anomaly detection and time series motif discovery.