Mining sensor data from complex systems

Vespier, U.

Chapter 8

Conclusions

In this thesis, we discussed data mining methods and algorithms for the unsupervised analysis of time series sensor data from complex physical systems. As complex systems are often affected by several phenomena at once, we developed techniques able to cope with these complexities, such as the presence of noisy measurements, multiple temporal scales and recurring patterns at play at the same time. We also presented a technique and a software tool to make the interactive visualization of massive time series datasets feasible.

We summarize our contributions below.

- In Chapter 4, we introduced a data mining method to discover the relevant temporal scales in a time series. We employ the scale-space theory and the Minimum Description Length principle to select the most relevant time series decomposition among the many possible by sub-diving the scale-space image. We introduced two different encoding schemes aimed at exploiting (making possible to compress) different properties of the data. We have shown how the presented methodology gives meaningful results in both artificial and real-world scenarios.
- A byproduct of the method is that it produces an optimal decomposition such that each component is represented according to its inherent complexity and does not present interferences from phenomena at other

temporal scales. We have shown an example of how these individual per-component representations may better serve tasks like classification, regression or association analysis than the input, mixed, time series.

- In Chapter 5, we introduced a method for the discovery of multi-scale motifs in time series data. Another typical property of complex systems is the presence of recurring phenomena at different temporal scales. Moreover, these recurring phenomena appear warped in time, more or less intense or both throughout the data. We employ the scale-space theory and Minimum Description Length as the base of our methodology. Differently from much of the existing literature, we employ a definition of motifs based on structural similarity, other than one based on point-wise comparisons, in order to account for warping and deformations. We propose a way to transform the structural representation of a motif into a symbolic string which allows for fast matching by means of suffix arrays. The effectiveness of the method is proved on sensor data from several applications, including InfraWatch data.

- In Chapter 6, we focused on the problem of identifying traffic activity events in strain measurements. The proposed solution is based on subsequence clustering, a technique known to be prone to undesired behaviors and whose outcome is strongly dependent on the kind of data it is applied to. In view of this, we studied SSC in relation to the features of the strain data, showing that only some of the documented pitfalls (i.e., multiple representations) occur in our case. To solve this, we introduced a context-aware distance measure between subsequences, which also takes the local neighborhood of a subsequence into account. Employing this *Snapping* distance measure, we showed that SSC by k-Means returns a correct modeling of the traffic events.

- In Chapter 7, we introduced the task of large time series visualization in the context of Exploratory Data Analysis (EDA). We proposed a storage scheme to hold sub-sampled versions of the original data in order to speed up data retrieval at different levels of resolution. Based on this storage technique, we presented a data retrieval mechanism

for visualization and VizTool, a software solution to support fast and interactive visualization of large time series data collected from sensors. VizTool was instrumental in discovering important properties and events in real-world sensor data such as InfraWatch, where it was helpful to discover dead sensors, re-calibration activities, correlation between temperature and strain response, differences of traffic activity between work days and weekend days.

One of the main themes of the thesis is how complex systems often exhibit diverse behaviors at different temporal scales. A general conclusion is that data mining methods should be able to cope with the multiple resolutions (scales) at the same time in order to fully understand the data at hand and extract useful information from it. This becomes more and more important as we advance in our ability to collect increasingly detailed data about the phenomena around us.

Throughout the thesis, we developed data mining methods aimed at coping with this challenge. An important conclusion is that the Minimum Description Length principle, paired with the scale-space construction, represents an effective way of discerning what is fundamental and what is not in the data while considering different scales of analysis. This same principle was instrumental in developing effective techniques to both detect the relevant scales and mine the patterns that occur.

The importance of looking at different scales is also connected to visualization. Complex systems show different phenomena at different resolution and interactive visualization can effectively help practitioners to focus on a specific resolution without being overwhelmed by the finer details.

One nice property of the methods we developed is that they do not require parameters. This is a result of our choice of employing the MDL principle and a model selection approach. However, we note that, in order for this approach to be effective in practice, it is important to define flexible encoding schemes that are able to capture the variability of the data at hand, possibly incorporating domain knowledge.

8.1 Future Work

Future work includes extending the presented methods to work in a streaming context. This would allow, for example, to discover new phenomena in a quasi-real time fashion or detect the presence of novel recurring patterns.

Another promising opportunity for future work is to explore how MDL and the scale-space theory could be used to design anomaly detection techniques at multiple scales. Anomaly detection is, in fact, an important task when dealing with the monitoring of complex systems, and MDL, other than as a model selection principle, could be employed to spot changes in the underlying data generation processes.

Finally, a relevant task linked to the goals of InfraWatch is the analysis of multivariate time series data in order to mine key performance indicators (KPIs) that behave in an approximately monotonic way. In fact, when looking for indicators such as the degradation of concrete structures, as in the case in InfraWatch, one approach is to look for subsets of sensors that, when combined through linear and non-linear operators, result in a monotonic time series. Such derived KPIs are clearly indicators of some underlying process that irreversibly moves in a certain direction, and degradation of the bridge is a good candidate for that. Using an approach called Equation Discovery [24], one might discover such monotonic functions.