Recent Trends in Learning of structured and non-standard data

Frank-Michael Schleif¹, Peter Tino¹ and Thomas Villmann²

1- University of Birmingham
School of Computer Science,
Edgbaston B15 2TT Birmingham,
United Kingdom.

2- University of Applied Sciences,
Department of Mathematics
Mittweida, Germany

Abstract. In many application domains data are not given in a classical vector space but occur in form of structural, sequential, relational characteristics or other non-standard formats. These data are often represented as graphs or by means of proximity matrices. Often these data sets are also huge and mathematically complicated to treat requesting for new efficient analysis algorithms which are the focus of this tutorial.

1 Introduction

Computational intelligence methods for the extraction of knowledge from large structured and non-standard data is becoming more and more important. Modern measurement systems e.g. in the life sciences and cheap storage devices facilitate the quick gathering of information at large scale. Considering projects like the human genome project or next generation sequencing, huge sets of DNA sequences are recorded which are inherently structural, waiting for analysis and knowledge extraction [4]. Many and large databases store proteins, phylogenetic trees, molecular graphs, time series data, hyper-textual and XML documents from the web all in more or less structured but clearly non-standard form often omitting a vector space representation. Some typical formats are depicted in Figure 1. Standard data sets typically consist of \(N\) samples in a real \(D\) di-

Fig. 1: Typical non-standard and structured data formats. Plot a) a graph representing relations and strength of proximity between objects Plot b) a time series and various encodings including Symbolic Aggregate approXimation. Plot c) a similarity matrix obtain from sequence scores.
mensional vector space. Non-standard and structured data as considered in the following may deviate from this assumption in different ways. Time series data, for example often consist of vectors with different lengths but in the same $D$ dimensional vector space. A similar problem is observed with sequential data where the $N$ sample are typically not given in a vector space but by means of a symbol sequence of different length on a common alphabet of symbols. Also different life science data like spectral measurements are not completely standard in the way that the dimensions imply structural information and are not independent. While for example for images a vectorial representation is common it may also be better to keep the natural matrix structure to preserve neighborhood relations. Similar also video or multivolume data e.g. for functional magnetic resonance imaging (fMRI) may be better represented in tensor form. Some data are also naturally represented by means of tree or graph structures also providing a source of non-standard and structured data sets. These variations form normal data sets can be approached in different ways most often by means of an adapted representation and or a specific model and learning algorithm.

Different dedicated algorithms are already proposed for formats like time-series, structured, sequence or matrix data [1, 2, 3, 4, 5] but still various challenges remain if it comes e.g. to large scale problems, the integration of auxiliary information, the handling of missing values or topics like semi-supervision. This tutorial provides a brief review and introduction to recent challenges and techniques in the field on structured- and non-standard data analysis.

2 Proximity matrix learning

A common strategy to process structured and non-standard data is to find a data representation such that more classical algorithms become accessible. Available similarity or dissimilarity measures for the specific formats can be used to encode the data such that a proximity matrix is obtained [6].

Considering e.g. the case of timeseries data, than such measures could be dynamic time warping or global alignment kernels [7]. For data like DNA or protein sequences (see Figure 2) alignment functions based on biological knowledge like the Smith-Waterman Algorithm or blast could be used [8].

A very generic measure is the Kolmogorov-Complexity which can be approximated by the normalized compression distance [9] often used to analyze textual documents. Given the obtained measure is metric, standard approaches like a kernel PCA, kernel k-Means, Support Vector Machine or Laplacian Eigenmap [5, 10] can be used. For non-metric measures [11] the situation is slightly more complicated because the obtained non-metric dissimilarities or indefinite kernels are less common and only few methods are available [12, 13, 14, 15, 16] or transformation methods or embedding methods are needed [6, 17].

The most common approaches to convert a non-metric similarity matrix $S$, as obtained e.g. from a sequence alignment, to a positive semi definite matrix are based on eigenvalue corrections like: flipping, clipping, vector-representation, shift correction. The underlying idea is to remove negative eigenvalues in the eigenspectrum of the matrix $S$. One may also try to learn an alternative psd
kernel representation with maximum alignment to the original non-psd kernel matrix \([18, 6, 19]\) or split the proximities into positive and negative eigenvalues as discussed in \([20, 21]\), both with high computational costs.

But the appropriate encoding of the data is not the only point to care about but also the number of points. Obviously for a quadratic proximity matrix \(S_{N \times N}\), with \(N\) as the number of samples larger data sets e.g. \(N > 10000\) become very challenging. Given the rank of the matrix is not to large efficient approximation strategies like the Nyström approximation \([22]\) or random approximation strategies \([23]\) can be used, which can also be used for indefinite proximity matrices \([24, 25]\).

Most of the aforementioned approaches rely on symmetric matrices but recently also some strategies for asymmetric analysis of proximity matrices were proposed \([26, 27]\).

### 3 Matrix completion and collaborative filtering

Non-standard data are also often given as an incomplete set of (weighted) relations between objects, which could be represented by a very sparse graph. Common examples are e.g. recommendation systems. The users of such a system (continuously) generate training data by ranking items (e.g. bought books) and other users may rank similar or the same items. The system tries to infer rules to predict items which are of interest for a user based on this knowledge-base most often employing meta information, like the fact that a specific set of items belongs to a common group. Matrix completion, collaborative filtering and low-rank matrix estimation are concepts used in this line \([28, 29, 30]\) and are very challenging as standard settings have been shown to be NP-hard \([31]\).

For matrix completion the data matrix \(Z\) is typically \(N \times M\) with \(M < N\). The rows can be e.g. customers and the columns refer to votes of articles the customers may have been bought. In general this matrix is very sparse because a customer bought only few of the \(M\) items. A typical sparsity is in the range of 99% zero entries e.g. for the famous Netflix database. The objective is now to fill
the missing values for an individual customer to predict its potential votes. To phrase this setting as a learning problem one assumes that $Z$ lies in a much lower dimensional manifold and can be represented e.g. as $Z \approx V_{N \times k} \cdot G_{k \times M}$ with $k \ll \min\{N, M\}$. $G$ may be considered to be a grouping matrix with a small number of groups and $g_{lj} \in G$ as the relative score of item $j$ in group $l$. The values $v_{il} \in V$ define the affinity of customer $i$ to the group $l$. Accordingly the modeled score for customer $i$ on item $j$ is the $\sum_{l=1}^{k} V_{il} \times G_{lj}$ of group affinities times group scores. This problem can be formulated as a constrained optimization problem and different variants were proposed to make it feasible e.g. for larger datasets.

Current work in this field tries to provide more specific analysis about minimal requirements to reconstruct a low rank matrix [32] or considers alternative norms and optimization schemes to obtained better results [33]. A very recent survey on Bayesian techniques for low rank matrix estimation was provided in [29].

4 Models for structured and non-standard data

Since already more than a decade learning and mining of structured and non-standard data has becoming more and more into the focus of the research community. Starting with early work around recursive neural networks [34], successful methods like the Self-Organizing-Map for structured data (SOM-SD) and variants [35, 36, 37] or very recent approaches for tree-structured data [38, 39, 40] have successfully tackled different data analysis problems for structured data as shown in [41]. An example how to generate a structural data representation model of a complex set of input data is depicted in Figure 3 (details in [38]). Recent work on representing biological sequence data by means of effective tree kernels can be found in [44]. An efficient strategy to integrate auxiliary information for spectral data in the model process is proposed in [45] whereas an example of non-trivial metric adaptation for microarray and spectral data is given in [46].

As another source of non-standard data and models one may also consider multiway array data as a generalization of matrices [47]. Such data typically occur in the line of medical fMRI and EEG measurements but also three dimensional video sequences or sensory data in the context of robotics are of this type. The corresponding tensor based algorithms can often better use the multilinear
structural properties of these datasets.

Time series data, as a specific form of non-standard data, have raised continuous interest in all times [1] but with new measurement systems (e.g. long term measurements in life sciences) and data sources in the Internet (e.g. click statistics, network logging data, high-frequency trading) [42] not only the amount of data has raised but also the dimensionality and the properties like varying sampling rate, sparsity and length have changed. These challenges motivated new dedicated methods like the Generalized Topographic Model Through Time (GTM-TT) which can effectively deal with multi-dimensional timeseries [48, 49] and was also extended by metric adaptation and supervision concepts [50, 51] and more recent approaches are based on reservoir computing and kernel techniques [52].

One prominent approach for large sets of time series data is Symbolic Aggregate approXimation (SAX) [53] which was generalized in [54] to take specific data properties into account. Also the approach proposed in [55] considers time series data but improved the learning of a regression model by employing privileged information about the data. Another recent approach linking to the former section was provided in [56] where multi-dimensional time series data are mapped into a parameter space of a hidden markov model employing a discriminative fisher kernel. A semi-supervised hidden random field based approach for timeseries was recently proposed in [57].

5 Conclusions

In this tutorial we briefly reviewed challenges and approaches common in the field of non-standard and structured data analysis. The more recent proposals in these domains focus on the analysis of large scale problems and the effective integration of meta information. The different strategies strongly depend on the chosen or given data representation. For matrix data the integration of meta information may be realized by learning an aligned matrix within an optimization problem [18]. Large scale problems for matrix data are address by matrix approximation approaches [22, 24, 23] or by learning algorithms which are also effective on few available proximity data like core set techniques coupled with probabilistic sampling [58, 59]. More recently also sparse probabilistic models where proposed which do not rely on metric proximity data [15]. Also for structural and sequence data large scale datasets remain challenging and approximations approaches like SAX are very popular [53]. But also here the integration of meta data by learning appropriate kernel representation [36, 52], also for streaming data [60], or by explicitly optimizing alignment function during training [61] or other metric adaptation approaches [46] show promising new directions. Also the explicit analysis of tensor data by means of dedicated methods [47] gets more interest.

Acknowledgment

This work was funded by a Marie Curie Intra-European Fellowship within the 7th European Community Framework Program (PIEF-GA-2012-327791).
References


