Semi-Supervised Structuring of Complex Data*

Marian-Andrei Rizoiu
ERIC Laboratory
University Lumière Lyon 2
Lyon, France
Marian-Andrei.Rizoiu@univ-lyon2.fr

Abstract

The objective of the thesis is to explore how complex data can be treated using unsupervised machine learning techniques, in which additional information is injected to guide the exploratory process. Starting from specific problems, our contributions take into account the different dimensions of the complex data: their nature (image, text), the additional information attached to the data (labels, structure, concept ontologies) and the temporal dimension. A special attention is given to data representation and how additional information can be leveraged to improve this representation.

1 The big picture

The general context of the research in this thesis lies at the intersection of **Complex Data Analysis** and **Semi-Supervised Clustering**. The research project behind the thesis was built incrementally through a dialectical relation between theory and practice. The research projects in which I was involved raised several precise problems, which usually dealt with handling complex data (heterogeneous data of different natures, *e.g.*, text, image) and embedding additional information into the learning process.

The different contributions of the thesis can be structured conceptually as show in Figure 1. The center points are translating data of different natures into a representation format understandable by machine learning algorithms and knowledge injection into the unsupervised learning algorithms and data representation constructors.

1.1 Dealing with Complex Data

Complex Data is difficult to process efficiently [Zighed *et al.*, 2009]. Its nature can be very diverse (*i.e.*, text, image or audio/video data). It is heterogeneous, as it may come from multiple sources. It is often temporal, as the evolution of different entities is recorded through time. Also, additional information and knowledge is attached to the data, under the form of user labels, structure of interconnected documents or knowledge bases.

*PhD prepared under the supervision of Stéphane Lallich and Julien Velcin, professors at the University Lumière Lyon 2.

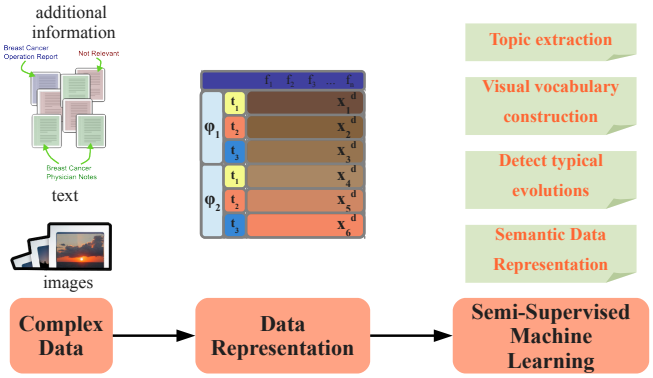

Figure 1: Conceptual organization of the work in the thesis.

A straightforward method for dealing with complex data of different natures is to transform them into a tabular numerical format and to apply classical Machine Learning algorithms. Most Data Mining algorithms were developed to use data in this format. The tabular format is a simple way to describe an instance as a measurement vector on a set of predefined features. Today's challenges lie in rendering the data into a common usable numeric format which succeeds in capturing the information present in the native format, and in efficiently using external information for both creating the format and improving the results of analysis.

1.2 Semi-supervised Clustering

Leveraging partial expert knowledge into clustering represents the domain of semi-supervised clustering. Unlike semi-supervised learning, where the accent is on dealing with missing data in supervised algorithms, semi-supervised clustering is used when the expert knowledge is in such low quantity that is would be impossible to apply supervised techniques. The expert knowledge is under the form of either class labels, or pairwise constraints, and it is used to guide the clustering process in the solutions space. Both the additional information that comes with the complex data and its temporal component can be modeled using semi-supervised clustering techniques.

2 Contributions

Detecting Typical Evolutions One of our driving interests is how to leverage the temporal information into clustering.

In [Rizoiu *et al.*, 2012], we detect typical evolutions of entities by proposing a new temporal-aware dissimilarity measure and a segmentation contiguity penalty function. We combine the spatio-temporal dimension of complex data with a semi-supervised clustering technique. We propose a novel time-driven constrained clustering algorithm, called TDCK-Means, which creates a partition of coherent clusters, both in the multidimensional space and in the temporal space.

Using Data Semantics to Improve Data Representation As Section 1.1 shows, treating data of different types (image, text) usually boils down to rendering the data into a Tabular Numeric Format and applying classical Machine Learning algorithms. We give a special importance to improving representation of the data in this format by using the underlying semantics of the dataset. We seek to construct, using unsupervised algorithms, new features that are more appropriate for describing the dataset and, at the same time, which are comprehensible for a human user. We propose in [Rizoiu *et al.*, 2013] two algorithms that construct the new features as conjunctions of the initial primitive features or their negations. The generated feature sets have reduced correlations between features and succeed in catching some of the hidden relations between individuals in a dataset.

Dealing with Text: topic extraction and evaluation Textual data can be rendered into the Tabular Numerical Format by using the "bag-of-words" representation. Once this representation set up, topics can be extracted and used at the Terms and Synonyms layers of the Ontology Learning Layer Cake as building blocks of constructing ontologies of concepts [Rizoiu and Velcin, 2011]. Topic extraction can benefit greatly from using additional information, under the form of ontologies of concepts. In [Musat *et al.*, 2011], we show how a concept ontology can be used to evaluate the topics extracted using graphical approaches (*e.g.*, LDA [Blei *et al.*, 2003]).

Improving Image Representation using Semi-supervised Visual Vocabulary Construction One of the most widely used way of changing the representation of images from the native format to the Tabular Numerical Format is the "bag-of-features" representation. We are interested in using expert knowledge, under the form of labels attached to the images, in the process of creating the numerical representation, through means of semi-supervised clustering. We propose two approaches: the first one is a tag-based visual vocabulary construction algorithm, while the second deals with filtering the background features from the object related features.

3 Conclusion. Current and future work.

The work in this thesis lies at the intersection of **Complex Data Analysis** and **Semi-Supervised Clustering**. We investigate how data of different natures can be treated, while considering the temporal dimension and the additional information that may come with the data. The domain of this thesis is vast and plenty of work still remains. Future work includes a better integration of the different proposed approaches and extending the type of data, the knowledge that can be used as additional information (*e.g.* processing video, using knowledge from the semantic web *etc.*).

3.1 Current work

We are presently working on extending and improving our current proposals. We are experimenting with an extension of our temporal clustering algorithm that infers a cluster graph structure simultaneously with the cluster construction. We are also interested in improving our feature construction algorithm, so that it uses the temporal dimension in addition to data semantics to improve data representation.

Another direction of our current work is to generalize the use of our proposals to other related use cases. For example, we apply our temporal clustering algorithm to the case of social networks. The idea is to detect temporally coherent user social roles in web forum discussion.

Applied work

The theoretical aspects of the thesis were doubled by a practical prototype production. The most prominent produced software is *CommentWatcher*, an open source tool aimed at analyzing discussions on web forums. Constructed as a web platform, *CommentWatcher* features automatic fetching of the forums using a versatile parser architecture, topic extraction from a selection of texts and a temporal visualization of extracted topics and the underlying social network of users. It is aimed at both the media watchers (it allows quick identification of important subjects in the forums and user interest) and the researches in social media (who can use to constitute temporal textual datasets).

References

[Blei *et al.*, 2003] David M. Blei, Andrew Y. Ng, and Michael I Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993–1022, 2003.

[Musat *et al.*, 2011] Claudiu Musat, Julien Velcin, Stefan Trausan-Matu, and Marian-Andrei Rizoiu. Improving topic evaluation using conceptual knowledge. In *IJCAI 2011*, volume 3, pages 1866–1871, 2011.

[Rizoiu and Velcin, 2011] Marian-Andrei Rizoiu and Julien Velcin. Topic extraction for ontology learning. In Wilson Wong, Wei Liu, and Mohammed Bennamoun, editors, *Ontology Learning and Knowledge Discovery Using the Web: Challenges and Recent Advances*, chapter 3, pages 38–61. 2011.

[Rizoiu *et al.*, 2012] Marian-Andrei Rizoiu, Julien Velcin, and Stéphane Lallich. Structuring typical evolutions using temporal-driven constrained clustering. In *ICTAI 2012*, pages 610–617, 2012.

[Rizoiu *et al.*, 2013] Marian-Andrei Rizoiu, Julien Velcin, and Stéphane Lallich. Unsupervised feature construction for improving data representation and semantics. *Journal of Intelligent Information Systems*, 2013.

[Zighed *et al.*, 2009] Djamel A. Zighed, Shusaku Tsumoto, Zbigniew W. Ras, and Hakim Hacid, editors. *Mining Complex Data*, volume 165. Springer, 2009.