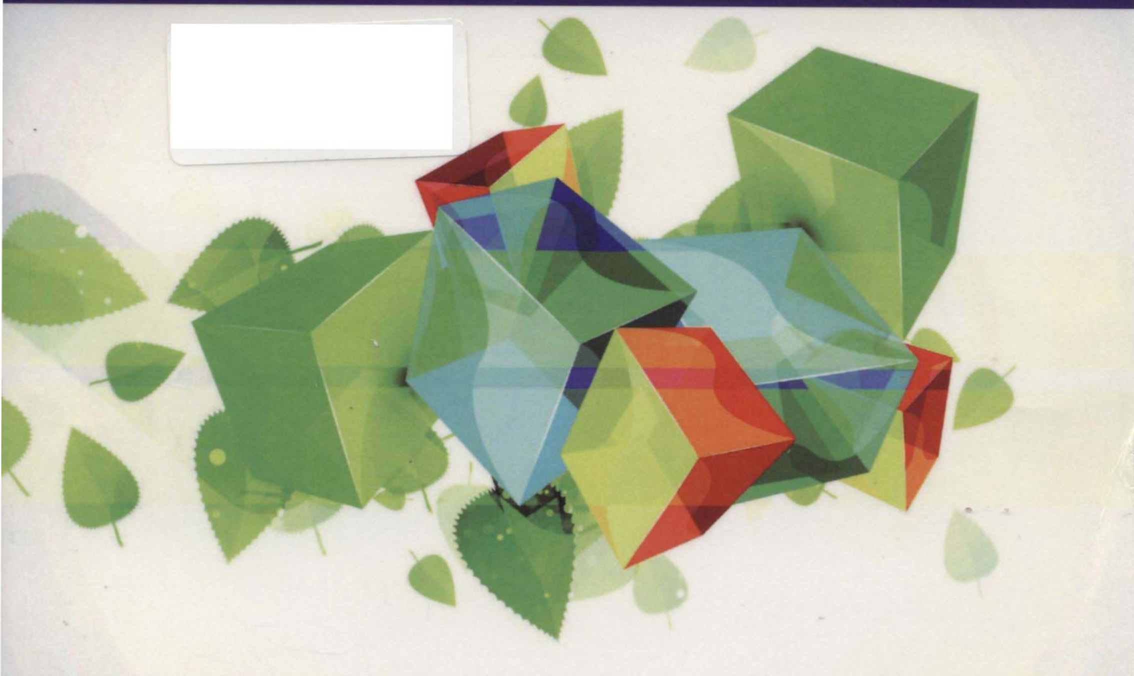

Giovanni Ponti

Advances in Mining Complex Data: Modeling and Clustering

LAP LAMBERT Academic Publishing

Giovanni Ponti

Advances in Mining Complex Data: Modeling and Clustering

LAP LAMBERT Academic Publishing

Impressum / Imprint
Bibliografische Information der Deutschen Nationalbibliothek: Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

Bibliographic information published by the Deutsche Nationalbibliothek: The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Coverbild / Cover image: www.ingimage.com

Verlag / Publisher:
LAP LAMBERT Academic Publishing
ist ein Imprint der / is a trademark of
AV Akademikerverlag GmbH & Co. KG
Heinrich-Böcking-Str. 6-8, 66121 Saarbrücken, Deutschland / Germany
Email: info@lap-publishing.com

Herstellung: siehe letzte Seite /
Printed at: see last page
ISBN: 978-3-659-30522-1

Zugl. / Approved by: Arcavacata di Rende (CS), University of Calabria, DEIS Dept., 2010

Giovanni Ponti

Advances in Mining Complex Data: Modeling and Clustering

Acknowledgments

It is a honor for me to be part of a research group where people work with devotion and passion in a friendly atmosphere. This environment has provided me with incitements and support during my studies.

The first acknowledgment goes to my advisor Prof. Sergio Greco. He was the first who believed in my capabilities and introduced me in the scientific research world. His suggestions were very precious and his intuitions opened my mind to many interesting research opportunities.

A special thanks to Andrea Tagarelli, who led me carefully to and through my research activities. His critical observations and his constructive remarks contributed to the growth of my research capabilities. He also revealed to be a friend who provided encouragements in difficulties. His presence is invaluable, a reference point in work and in life.

I also acknowledge Prof. Luigi Palopoli, who directed the Ph.D. course. He was always interested in the evolution of my activities and his suggestions were valuable to enrich my studies.

The majority of the topics shown in this thesis are result of research activities conducted with my room mate, Francesco Gullo. Such results could not be obtained without his contribution. We worked together sharing opinions and providing precious remarks each other. Apart from work, we shared many moments of friendship. For these reasons, he deserves a great thanks.

There were many people at DEIS who stood by me during these years. All of them gave me an important contribution to improve my knowledge with helpful feedback, discussions, and complementary points of view. In particular, I would thank Bettina Fazzinga, Sergio Flesca, Filippo Furfaro, Massimiliano Mazzeo, Cristian Molinaro, Francesco Parisi, Andrea Pugliese, Francesca Spezzano, and Ester Zumpano for work and joy moments spent together. I also thank Mario Cannataro, Pierangelo Veltri, and Giuseppe Tradigo from UNICZ, who provided me a valuable support in several issues of my research experience. I acknowledge Giovanni

Costabile and Gabriele Gigliotti for their help in bureaucratic affairs.

Apart from the work environment, I would thank people with whom I spent my leisure time. For this reason, a great thanks goes to all my friends of "Post Cresima". I get known them since the youth and we grew up together, sharing lots of joyful and spiritual moments.

An invaluable thanks goes to my parents Luigi and Mena, as well as my sister Barbara. They always supported me in pursuing my objectives. A whole life would not be enough to express all my gratitude to them.

Last but not least, my deepest thanks is for Mari, my better half. She is the most important person in my life, and her love gave me the strength to overcome difficulties I faced during these years. Also, discussions and moments spent together were very important for my overall growth. My life is better since we met.

to Mari

Preface

In the last years, there has been a great production of data that come from different application contexts. However, although technological progress provides several facilities to digitally encode any type of event, it is important to define a suitable representation model which underlies the main characteristics of the data. This aspect is particularly relevant in fields and contexts where data to be archived can not be represented in a fix structured scheme, or that can not be described by simple numerical values. We hereinafter refer to these data with the term *complex data*.

Although it is important define ad-hoc representation models for complex data, it is also crucial to have analysis systems and data exploration techniques. Analysts and system users need new instruments that support them in the extraction of patterns and relations hidden in the data. The entire process that aims to extract useful information and knowledge starting from raw data takes the name of *Knowledge Discovery in Databases* (KDD). It starts from raw data and consists in a set of specific phases that are able to transform and manage data to produce models and knowledge. There have been many knowledge extraction techniques for traditional structured data, but they are not suitable to handle complex data.

Investigating and solving representation problems for complex data and defining proper algorithms and techniques to extract models, patterns and new information from such data in an effective and efficient way are the main challenges which this thesis aims to face. In particular, two main aspects related to complex data management have been investigated, that are the way in which complex data can be modeled (i.e., *data modeling*), and the way in which homogeneous groups within complex data can be identified (i.e., *data clustering*). The application contexts that have been objective of such studies are *time series data*, *uncertain data*, *text data*, and *biomedical data*.

It is possible to illustrate research contributions of this thesis by dividing them into four main parts, each of which concerns with one specific area and data type:

Time Series — A time series representation model has been developed, which is conceived to support accurate and fast similarity detection. This model is called *Derivative time series Segment Approximation (DSA)*, as it achieves a concise yet feature-rich time series representation by combining the notions of *derivative estimation*, *segmentation* and *segment approximation*.

Uncertain Data — Research in uncertain data mining went into two directions. In a first phase, a new proposal for partitional clustering has been defined by introducing the *Uncertain K-medoids (UK-medoids)* algorithm. This approach provides a more accurate way to handle uncertain objects in a clustering task, since a cluster representative is an uncertain object itself (and not a deterministic one). In addition, efficiency issue has been addressed by defining a distance function between uncertain objects that can be calculated offline once per dataset.

In a second phase, research activities aimed to investigate issues related to hierarchical clustering of uncertain data. Therefore, an agglomerative centroid-based linkage hierarchical clustering framework for uncertain data (*U-AHC*) has been proposed. The key point lies in equipping such scheme with a more accurate distance measure for uncertain objects. Indeed, it has been resorted to *information theory* field to find a measure able to compare probability distributions of uncertain objects used to model uncertainty.

Text Data — Research results on *text data* can be summarized in two main contributions. The first one regards clustering of *multi-topic documents*, and a framework for hard clustering of documents according to their mixtures of topics has been proposed. Documents are assumed to be modeled by a generative process, which provides a mixture of probability mass functions (pmfs) to model the topics that are discussed within any specific document. The framework combines the expressiveness of generative models for document representation with a properly chosen information-theoretic distance measure to group the documents.

The second proposal concerns *distributional clustering of XML documents*, focusing on a the development of a distributed framework for efficiently clustering XML documents. The distributed environment consists of a peer-to-peer network where each node in the network has access to a portion of the whole document collection and communicates with all the other nodes to perform a clustering task in a collaborative fashion. The proposed framework is based on modeling and clustering XML documents by structure and content. Indeed, XML documents are transformed into *transactional data* based on the notion of *tree tuple*. The framework is based on the well-known paradigm of *centroid-based partitional clustering* to conceive the distributed, transactional clustering algorithm.

Biomedical Data — Research results on time series and uncertain data have been involved to support effective and efficient biomedical data management. The focus regarded both proteomics and genomics, investigating Mass Spectrometry (MS) data and microarray data. In the specific, a *Mass Spectrometry Data Analysis (MaSDA)* system has been defined. The key idea consists in exploiting temporal information implicitly contained in MS data and model such data as time series. The major advantages of this solution are the dimensionality and the noise reduction. As regards micrarray data, U-AHC has been employed to perform clustering of microarray data with probe-level uncertainty. A strategy to model probe-level uncertainty has been defined, together with a hierarchical clustering scheme for analyzing such data. This approach performs a gene-based clustering to discover clustering solutions that are well-suited to capture the underlying gene-based patterns of microarray data.

The effectiveness and the efficiency of the proposed techniques in clustering complex data are demonstrated by performing intense and exhaustive experiments, in which such proposals are extensively compared with the main state-of-the-art competitors.

Giovanni Ponti

Contents