Advances in Computer Vision and Pattern Recognition

Marcello Pelillo *Editor*

Similarity-Based Pattern Analysis and Recognition

Springer

Advances in Computer Vision and Pattern Recognition

For further volumes:
www.springer.com/series/4205

Marcello Pelillo
Editor

Similarity-Based Pattern Analysis and Recognition

Springer

Editor
Marcello Pelillo
DAIS
Ca' Foscari University
Venice, Italy

Series Editors
Sameer Singh
Rail Vision Europe Ltd.
Castle Donington
Leicestershire, UK

Sing Bing Kang
Interactive Visual Media Group
Microsoft Research
Redmond, WA, USA

ISSN 2191-6586 ISSN 2191-6594 (electronic)
Advances in Computer Vision and Pattern Recognition
ISBN 978-1-4471-5627-7 ISBN 978-1-4471-5628-4 (eBook)
DOI 10.1007/978-1-4471-5628-4
Springer London Heidelberg New York Dordrecht

Library of Congress Control Number: 2013955585

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

For my parents, who made it possible

“Surely there is nothing more basic to thought and language than our sense of similarity. [. . .]
And every reasonable expectation depends on resemblance of circumstances, together with our tendency to expect similar causes to have similar effects.”

Willard V.O. Quine

Foreword

The SIMBAD project was a Future and Emerging Technologies (FET) project funded by the European Commission between 2008 and 2011. It brought together an extraordinary group of talented researchers with a broad spectrum of different perspectives on the central theme of using non-Euclidean similarity functions as the basis for learning. This approach was in contrast with the use of kernel functions that had become the de facto standard at the time of the project's launch in 2008.

The SIMBAD project took a broad view of the problem of so-called non-Euclidean learning: analysing the extent to which this was essential in a particular problem, developing alternative learning strategies that could successfully learn from non-Euclidean similarity functions, developing methods of learning Euclidean representations from probabilistic models and similarity data, and so on. These approaches were not studied just in the abstract but rather were grounded in a series of concrete problems from application domains where it was known or suspected that the Euclidean assumption was unrealistic.

The number and depth of the papers that arose from this research agenda was very impressive, with significant innovations made on all of the fronts listed above. However, the research was not merely a shotgun attack on several divergent fronts, but rather represented the coherent development of the leitmotiv of the project: the use of similarity functions in learning.

Given the breadth of the reach and impact of the research, the project reviewers were fearful that this coherence might be lost in the variety of journals, conferences, and particular problems considered, hence risking that the main message become lost in the plethora of individual results.

It was therefore proposed that a book bringing together the themes of the project and its main results could help champion and communicate the SIMBAD message in one coherent volume. This carefully constructed book is the result of that proposal. It is a distillation of the main themes and results of the project into an accessible and cross-referenced volume. For those interested in learning about the potential and importance of learning from similarity functions, this work is undoubtedly the

key reference from which to begin their study and it is likely to remain so for many years to come.

Virginia Water
June 2013

John Shawe-Taylor

Preface

This book provides a thorough description of a selection of results achieved within SIMBAD, an EU FP7 project which represents the first systematic attempt at bringing to full maturation a paradigm shift that is just emerging within the pattern recognition and machine learning domains, where researchers are becoming increasingly aware of the importance of similarity information *per se*, as opposed to the classical (feature-based) approach.

SIMBAD started in April 2008 and ended in September 2011, and involved the following six partners:

- University of Venice, Italy (*scientific coordinator*)
- University of York, UK
- Delft University of Technology, The Netherlands
- Instituto Superior Tecnico, Lisbon, Portugal
- ETH Zurich, Switzerland
- University of Verona, Italy.

The very end of the project marked also the launch of the SIMBAD workshop series

http://www.dsi.unive.it/~simbad

whose first edition was held in Venice, in September 2011, in conjunction with the project's final review meeting. These biennial workshops aim to consolidate and promote research efforts in this area and to provide and informal discussion forum for researchers and practitioners.

Within the SIMBAD project we undertook a thorough study of several aspects of purely similarity-based pattern analysis and recognition methods, from the theoretical, computational, and applicative perspective. We covered a wide range of problems and perspectives. We considered both supervised and unsupervised learning paradigms, generative and discriminative models, and our interest ranged from purely theoretical problems to real-world practical applications. The chapters collected in this book aim to provide a coherent overview of our main achievements and to serve as a starting point for graduate students and researchers interested in

this important, yet diverse subject. More details on the project's activities can be found on our website

http://simbad-fp7.eu

and in the published papers referenced in this book.

A project like SIMBAD could not have been done without the help and support of many people and institutions, and it is a pleasure to take this opportunity to express my gratitude to them. In the first place, I'd like to acknowledge the Future and Emerging Technology (FET) Programme of the 7th Framework Programme for Research of the European Commission which funded the SIMBAD project, and I am very grateful to our project officer, Teresa De Martino, and to the reviewers, Georgios Sakas, Christoph Schnörr and John Shawe-Taylor, whose insightful suggestions and constant encouragement have been instrumental to make SIMBAD a better project.

It has been my good fortune to collaborate for almost four years with a fantastic group of people, whose genuine enthusiasm and exceptional professional competence made SIMBAD a unique, intellectually stimulating experience. In particular, I'm grateful to my fellow principal investigators who coordinated the activities of the various research units: Joachim Buhmann, Bob Duin, Mario Figueiredo, Edwin Hancock, and Vittorio Murino; to their deputies: Manuele Bicego, Umberto Castellani, Ana Fred, Marco Loog, Volker Roth, and Richard Wilson; and to all PhD students and postdocs who have worked within the project.

In Venice, I've been helped by many people in my group, and I'd like to thank them all for their support. In particular, I wish to thank Andrea Torsello for the assistance he gave me at various stages of the project, and Veronica Giove for her valuable work concerning all administrative aspects. Special thanks are due to Samuel Rota Bulò for his constant support throughout the project and for helping me assemble this book.

I'd like to thank the editorial staff at Springer, in particular Wayne Wheeler for supporting the idea of publishing this book, and Simon Rees for his advice throughout the production of the volume and for gently tolerating my procrastinations.

My deepest gratitude, however, go to my wife, Rosanna, and my children, Claudia and Valerio, without whose endless patience and understanding the SIMBAD project, and hence this book, would have not seen the light.

Venice
July 2013

Marcello Pelillo

Contents

Chapter 1
Introduction: The SIMBAD Project

Marcello Pelillo

Abstract This introductory chapter describes the SIMBAD project, which represents the first systematic attempt at bringing to full maturation a paradigm shift that is just emerging within the pattern recognition and machine learning domains, where researchers are becoming increasingly aware of the importance of similarity information per se, as opposed to the classical (feature-based) approach.

1.1 Motivations

The challenge of automatic pattern analysis and recognition (or machine learning) is to develop computational methods which learn, from examples, to distinguish among a number of classes, with a view to endow artificial systems with the ability to improve their own performance in the light of new external stimuli. This ability is widely recognized to be instrumental in building next-generation artificial cognitive systems (ACSs) which, as opposed to traditional machine or computer systems, can be characterized "as systems which cope with novel or indeterminate situations, which aim to achieve general goals as opposed to solving specific problems, and which integrate capabilities normally associated with people or animals."[1] The socio-economic implications of this scientific endeavor are enormous, as ACSs will have applications in a wide variety of real-world scenarios ranging from industrial manufacturing to vehicle control and traffic safety, to remote and on-site (environmental) sensing and monitoring, and to medical diagnostics and therapeutics.

As a matter of fact, despite their technological applications, pattern recognition and machine learning can arguably be considered as a modern-day incarnation of an endeavor which has challenged mankind since antiquity. Fundamental questions pertaining to categorization, abstraction, generalization, induction, etc. have, in fact, been on the agenda of mainstream philosophy, under different names and guises,

[1] From: *Artificial Cognitive Systems in FP7: A Report on Expert Consultations for the EU Seventh Framework Programme 2007–2013 for Research and Technology Development*.

M. Pelillo (✉)
DAIS, Università Ca' Foscari, Venice, Italy
e-mail: pelillo@dais.unive.it

M. Pelillo (ed.), *Similarity-Based Pattern Analysis and Recognition*,
Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-5628-4_1, © Springer-Verlag London 2013

since its inception. Indeed, as pointed out in [7], the very foundations of pattern recognition can be traced back to Aristotle and his mentor Plato who were among the firsts to distinguish between an "essential property" from an "accidental property" of an object, so that the whole field of pattern recognition can naturally be cast as the problem of finding such essential properties of a category. As Watanabe put it [20, p. 21]: "whether we like it or not, under all works of pattern recognition lies tacitly the Aristotelian view that the world consists of a discrete number of self-identical objects provided with, other than fleeting accidental properties, a number of fixed or very slowly changing attributes. Some of these attributes, which may be called 'features,' determine the class to which the object belongs." Accordingly, the goal of a pattern recognition algorithm is to discern the essences of a category, or to "carve the nature at its joints." In philosophy, this view is known as *essentialism* and has contributed to shape mainstream machine learning research in a such a way that it seems legitimate to speak about an essentialist paradigm.

During the nineteenth and the twentieth centuries, the essentialist world-view was subject to a massive assault from several quarters and it became increasingly regarded as an impediment to scientific progress. Strikingly enough, this conclusion was arrived at independently in at least three different disciplines, namely physics, biology, and psychology. In physics, anti-essentialist positions were held (among others) by Mach, Duhem, Poincaré, and in the late 1920s Bridgman, influenced by Einstein's achievements, put forcefully forward the notion of operational definitions precisely to avoid the troubles associated with attempting to define things in terms of some intrinsic essence [4]. For example, the (special) theory of relativity can be viewed as the introduction of operational definitions for simultaneity of events and of distance, and in quantum mechanics the notion of operational definitions is closely related to the idea of observables. This point was vigorously defended by Popper [15], who developed his own form of anti-essentialism and argued that modern science (and, in particular, physics) was able to make real progress only when it abandoned altogether the pretension of making essentialist assertions, and turned away from "what-is" questions of Aristotelian-scholastic flavor.

In biology, the publication of Darwin's Origin of Species in 1859 had a devastating effect on the then dominating paradigm based on the static, Aristotelian view of species, and shattered 2000 years of research which culminated in the monumental Linnaean system of taxonomic classification. According to Mayr [14], essentialism "dominated the thinking of the western world to a degree that is still not yet fully appreciated by the historians of ideas. [...] It took more than two thousand years for biology, under the influence of Darwin, to escape the paralyzing grip of essentialism."

More recently, motivated by totally different considerations, cognitive scientists have come to a similar discontent towards essentialist explanations. Indeed, it has become increasingly clear that the classical essentialist, feature-based approach to categorization is too restrictive to be able to characterize the intricacies and the multifaceted nature of real-world categories. This culminated in the 1970s in Rosch's now classical "prototype theory" which is generally recognized as having revolutionized the study of categorization within experimental psychology; see [13] for an

extensive account, and the recent paper by von Luxburg et al. [19] for a machine learning perspective.

Nowadays, anti-essentialist positions are associated with various philosophical movements including pragmatism, existentialism, decostructionism, etc., and is also maintained in mathematics by the adherents of the structuralist movement, a view which goes back to Dedekind, Hilbert and Poincaré, whose basic tenet is that "in mathematics the primary subject-matter is not the individual mathematical objects but rather the structures in which they are arranged" [16, p. 201]. Basically, for an anti-essentialist what really matters is relations, not essences. The influential American philosopher Richard Rorty nicely sums up this "panrelationalist" view with the suggestion that there are "relations all the way down, all the way up, and all the way out in every direction: you never reach something which is not just one more nexus of relations" [17]. As an aside, we note that a similar dissatisfaction with the essentialist approach can also be found in modern link-oriented approaches to network analysis [8, 12].

Now, it is natural to ask: What is the current state of affairs in pattern recognition and machine learning? As mentioned above, the fields have been dominated since their inception by the notion of "essential" properties (i.e., features) and traces of essentialism can also be found, to varying degrees, in modern approaches which try to avoid the direct use of features (e.g., kernel methods). This essentialist attitude has had two major consequences which have greatly contributed to shape the fields in the past few decades. On the one hand, it has led the community to focus mainly on feature-vector representations. Here, each object is described in terms of a vector of numerical attributes and is therefore mapped to a point in a Euclidean (geometric) vector space, so that the distances between the points reflect the observed (dis)similarities between the respective objects. On the other hand, this has led researchers to maintain a reductionist position, whereby objects are seen in isolation and which therefore tends to overlook the role of relational, or contextual, information.

Feature-vector representations are indeed extremely attractive because geometric spaces offer powerful analytical as well as computational tools that are simply not available in other representations. In fact, classical pattern recognition methods are tightly related to geometrical concepts and numerous powerful tools have been developed during the last few decades, starting from linear discriminant analysis in the 1920s, to perceptrons in the 1960s, to kernel machines in the 1990s. However, there are numerous application domains where either it is not possible to find satisfactory features or they are inefficient for learning purposes. This modeling difficulty typically occurs in cases when experts cannot define features in a straightforward way (e.g., protein descriptors vs. alignments), when data are high dimensional (e.g., images), when features consist of both numerical and categorical variables (e.g., person data, like weight, sex, eye color, etc.), and in the presence of missing or inhomogeneous data. But, probably, this situation arises most commonly when objects are described in terms of structural properties, such as parts and relations between parts, as is the case in shape recognition [3]. This led in 1960s to the development of the structural pattern recognition approach, which uses symbolic

data structures, such as strings, trees, and graphs for the representation of individual patterns, thereby, reformulating the recognition problem as a pattern-matching problem.

It is clearly open to discussion to what extent the lesson learnt from the historical development of other disciplines applies to machine learning and pattern recognition, but it looks at least like that today's research in these areas is showing an increasing propensity towards anti-essentialist/relational approaches. Indeed, in the last few years, interest around purely similarity-based techniques has grown considerably. For example, within the supervised learning paradigm (where expert-labeled training data is assumed to be available) the now famous "kernel trick" shifts the focus from the choice of an appropriate set of features to the choice of a suitable kernel, which is related to object similarities. However, this shift of focus is only partial as the classical interpretation of the notion of a kernel is that it provides an implicit transformation of the feature space rather than a purely similarity-based representation. Analogously, in the unsupervised domain, there has been an increasing interest around pairwise algorithms, such as spectral and graph-theoretic clustering methods, which avoid the use of features altogether. Other attempts include Balcan et al.'s theory of learning with similarity functions [2], and the so-called collective classification approaches, which are reminiscent of relaxation labeling and similar ideas developed in computer vision back in the 1980s (see, e.g., [18] and references therein).

Despite its potential, however, presently the similarity-based approach is far from seriously challenging the traditional paradigm. This is due mainly to the sporadicity and heterogeneity of the techniques proposed so far and the lack of a unifying perspective. On the other hand, classical approaches are inherently unable to deal satisfactorily with the complexity and richness arising in many real-world situations. This state of affairs hinders the application of machine learning techniques to a whole variety of relevant, real-world problems.

The main problem with purely similarity-based approaches is that, by departing from vector-space representations, one is confronted with the challenging problem of dealing with (dis)similarities that do not necessarily possess the Euclidean behavior [2] or not even obey the requirements of a metric. The lack of the Euclidean and/or metric properties undermines the very foundations of traditional pattern recognition theories and algorithms, and poses totally new theoretical/computational questions and challenges. In fact, this situation arises frequently in practice. For example, non-Euclidean or non-metric (dis)similarity measures are naturally derived when images, shapes or sequences are aligned in a template matching process. In computer vision, non-metric measures are preferred in the presence of partially occluded objects [27]. Other non-metric examples include pairwise structural alignments of proteins that focus on local similarity [5], variants of Hausdorff distance [18],

[2] A set of distances D is said to be *Euclidean* (or *geometric*) if there exists a configuration of points in some Euclidean space whose interpoint distances are given by D. In the sequel, the terms *geometric* and *Euclidean* will be used interchangeably. The term *(geo)metric* is an abbreviation to indicate the case of a distance that satisfies either the Euclidean or the metric properties.

normalized edit-distances [5], and also some probabilistic measures such as the Kullback–Leibler divergence. As argued in [27], the violation of the metric properties is often not an artifact of poor choice of features or algorithms, and it is inherent in the problem of robust matching when different parts of objects (shapes) are matched to different images. The same argument may hold for any type of local alignments. Corrections or simplifications may therefore destroy essential information.

In summary, there is an urgent need to bring to full maturation a paradigm shift that is just emerging within the pattern recognition and machine learning domains, where researchers are becoming increasingly aware of the importance of similarity information per se, as opposed to the classical feature-based (or vectorial) approach. Indeed, the notion of similarity (which appears under different names such as proximity, resemblance, and psychological distance) has long been recognized to lie at the very heart of human cognitive processes and can be considered as a connection between perception and higher-level knowledge, a crucial factor in the process of human recognition and categorization [9, 10].

1.2 The Structure of SIMBAD

SIMBAD represented the first systematic attempt towards the goal alluded to above. Within the project, we undertook a thorough study of several aspects of similarity-based pattern analysis and recognition methods, from the theoretical, algorithmic, and applicative perspective, with a view to substantially advance the state of the art in the field and contribute towards the long-term goal of organizing this emerging field into a more coherent whole.

We focused on two main themes, which basically correspond to the two fundamental questions that arise when abandoning the realm of feature-vector representations, namely:

1. How can one *obtain* suitable similarity information from object representations that are more powerful than, or simply different from, the vectorial?
2. How can one *use* similarity information in order to perform learning and classification tasks?

Although the two issues are clearly interrelated, it is advantageous to keep them apart as this allows one to separate the similarity generation process (a data modeling issue) from the learning and classification processes (a task modeling issue). According to this perspective, the very notion of similarity becomes the pivot of non-vectorial pattern recognition in much the same way as the notion of feature-vector plays the role of the pivot in the classical (geometric) paradigm. This results in a useful modularity, which means that all interactions between the object representation and the learning algorithm are mediated by the similarities, which is where the domain knowledge comes into the scene.

An important part of the project concerned the application of the developed techniques. To this end, we focused mainly on biomedical problems, which lend themselves particularly well to similarity-based approaches. Specifically, we applied the new methods developed within the project to inference tasks in the field of medical image analysis, i.e., to Tissue Micro Array (TMA) analysis and to Magnetic Resonance (MR) brain imaging.

Accordingly, the project (and hence this book) was structured around the following strands:

- Foundational issues
- Deriving similarities for non-vectorial data
- Embedding and beyond
- Applications

which we now briefly describe.

1.2.1 Foundational Issues

One of the first objectives within SIMBAD was to explore the causes and origins of non-Euclidean (dis)similarity measures and how they influence the performance of classical classification algorithms. In particular, we distinguished between the situation where the informational content associated with the violation of the geometric properties is limited, or is simply an artifact of the measurement process, and that where this is not the case. This distinction is important as, depending on the actual situation, two different strategies can be pursued: the first attempts to impose geometricity by somehow transforming or re-interpreting the similarity data, the second does not and works directly on the original similarities. Chapter 2 provides a comprehensive summary of our findings. It also discusses several techniques to convert non-Euclidean data into Euclidean and provides real-world examples which show that the non-geometric part of the data might be essential for building good classifiers.

A second line of investigation within this strand concerned fundamental questions pertaining to the very nature of the pattern recognition endeavor. Indeed, the search for patterns in data requires a mathematical definition of structure and a comparison function to rank different structures, thereby providing insights into the invariances in the problem class at hand. Motivated by an analogy between communication and learning, Chap. 3 describes an information-theoretic perspective to the problem and attempts to address the question of model selection and validation or, in other words, the tradeoff between informativeness and robustness. According to the proposed view, the notion of a pattern is interpreted as an element of an interpretation space (the "hypothesis class") endowed with a "natural" neighborhood system, or topology. By generalizing Shannon's random coding concept, the framework is able to determine which hypotheses are statistically indistinguishable due to measurement noise and how much we have to coarsen the hypothesis

class. The framework is thought to be applicable to more general questions arising in computer science concerning algorithm evaluation as well as (robust) algorithm design.

1.2.2 Deriving Similarities for Non-vectorial Data

The goal here was to develop suitable similarity measures for non-vectorial data. We focused primarily on structured data (e.g., strings, graphs, etc.), because of their expressive power and ubiquity, and on geometric measures as they allow one to employ the whole arsenal of powerful techniques available in the geometric pattern recognition literature. We pursued our goal by developing suitable kernels, which are known to be in correspondence with geometric (dis)similarities and considered in particular information-theoretic kernels. These are based on the assumption that the objects of interest are generated by some probabilistic mechanism (a source, in information/coding theoretic terms) and then proceed by defining (dis)similarity measures or kernels between (or among) models of these probabilistic sources. Chapter 4 reviews a recent approach which exploits the probabilistic nature of the so-called generative embeddings, by using information-theoretic kernels defined on probability distributions. This leads to a new class of hybrid generative/discriminative methods for learning classifiers whose effectiveness has been tested on two medical applications (see also Chaps. 9 and 10).

An alternative to this "kernel tailoring" approach consists in learning good similarities directly from training data. Within SIMBAD we investigated a strategy based on the evidence accumulation clustering paradigm, which aims to combine the results of multiple clusterings into a single data partition by viewing each clustering result as an independent evidence of data organization. Chapter 5 describes an approach which exploits the duality of similarity-based and probabilistic interpretations of the learned co-association matrix in order to produce robust and informative consensus solutions. This leads to two clustering methods: a "hard" method which explores embeddings over learned pairwise associations, and a unified probabilistic approach that we called PEACE (Probabilistic Evidence Accumulation for Clustering Ensembles), leading to soft assignments of objects to clusters.

1.2.3 Embedding and Beyond

Within this research strand, we aimed at developing computational models that do not depend on the actual object representation and rely only on (available) similarity information. As pointed out above, the analysis carried out in Chap. 2 suggests two complementary approaches. On the one hand, when the information content of non-geometricity is limited or simply caused by measurement errors, it is a plausible strategy to perform some correction on the similarity data (or finding an alternative

vectorial representation) in an attempt to impose geometricity, and then use conventional geometric techniques. On the other hand, when the information content of non-geometricity is relevant, one needs brand new tools, as standard approaches would not work in this case.

The former approach is known as "embedding," which is a well-established technique for vector-based representations, and is the subject of Chaps. 6 and 7. In particular, Chap. 6 focuses on two contrasting approaches to the problem. In the first part, it describes spectral methods for embedding structured data such as weighted graphs in a geometrically meaningful way. The resulting embeddings are then used to construct generative models for graph structure. To this end, the chapter explores the idea of "spherical" embedding, whereby data is embedded onto the surface of sphere of optimal radius. Instead of approximating the original (dis)similarities by Euclidean distances, the second approach tries to preserve the underlying group structure of the data. Within this context, the second part of Chap. 6 shows that a polynomial characterization derived from the Ihara zeta function leads to an embedding of hypergraphs which captures interesting structural properties.

Chapter 7 also focuses on these "structure-preserving" embeddings and restricts the discussion to the case of partition-based clustering problems. It is shown that a classical pairwise clustering cost function possesses an interesting shift-invariance property which amounts to saying that the choice of a partition is not influenced by additive constant shifts in the off-diagonal elements of the affinity matrix. An approximate version of this property is shown to hold in a more general probabilistic setting which is capable of selecting the number of clusters in a data-adaptive way. These findings raise intriguing questions concerning the role of structure-preserving embedding in the context of a theory of similarity-based pattern recognition.

When there is significant information content in the non-(geo)metricity of the data one has to resort to algorithms that work directly on the original similarity function. To this end, Chap. 8 describes an approach based on game theory which is shown to offer an elegant and powerful conceptual framework that serves well our purpose. The main point made by game theorists is to shift the emphasis from optimality criteria to equilibrium conditions, namely to the search of a balance among multiple interacting forces. Interestingly, the development of evolutionary game theory in the late 1970s offered a dynamical systems perspective, an element which was totally missing in the traditional formulation. From our perspective, one of the main attractive features of game theory is that it imposes no restriction whatsoever on the structure of the similarity function. Chapter 8 describes our attempts at formulating classical pattern recognition problems from a purely game-theoretic perspective. In particular, the chapter focuses on data clustering and structural matching and discusses some successful computer vision applications.

1.2.4 Applications

Pattern recognition and machine learning are essentially application-oriented fields with well-established validation techniques. These were used to quantitatively eval-

uate the success of the proposed research on large-scale applications with clear societal impact. In particular, within SIMBAD we devoted substantial effort towards tackling two large-scale biomedical imaging applications. With the direct involvement of leading pathologists and neuroscientists from the University Hospital Zurich and the Verona–Udine Brain Imaging and Neuropsychology Program, we contributed towards the concrete objective of providing effective, advanced techniques to assist in the diagnosis of renal cell carcinoma, one of the ten most frequent malignancies in Western countries, as well as of major psychoses such as schizophrenia and bipolar disorders. The results of our research are summarized in Chaps. 9 and 10, respectively. These problems are not amenable to be tackled with traditional machine learning techniques due to the difficulty of deriving suitable feature-based descriptions. For instance, image segmentation and shape alignment problems often produce non-(geo)metric dissimilarity data in both application domains, a feature which is indeed present in many other biomedical problems.

1.3 Conclusion and Outlook

There is an increasing awareness of the importance of similarity-based approaches to pattern recognition and machine learning, and research in this area has gone past the proof-of-concept phase and is now spreading rapidly. In fact, traditional feature-based techniques are felt as inherently unable to deal satisfactorily with the complexity and richness arising in many real-world situations, thereby hindering the application of machine learning techniques to a whole variety of relevant, real-world problems. Hence, in general, progress in similarity-based approaches will surely be beneficial for machine learning as a whole and, consequently, for the long-term enterprise of building intelligent systems.

We do believe that SIMBAD has contributed substantially towards the advancement of the state of the art in this area. In fact, we have introduced fresh perspectives to old problems, we have provided a thorough analysis of foundational issues, and we have demonstrated the applicability of our methodologies in real-world applications. In conclusion, we went far beyond our original expectations. Of course, we think there is room for improvement. In this respect, it might probably be useful to involve people from "external" fields such as cognitive psychology and/or algorithmics, thereby making the research more interdisciplinary. Also, as a matter of future work, there are promising application areas, such as chemometrics, bioinformatics, social network analysis, etc., which would certainly benefit from the work done within the project. We do hope that the availability into a single coherent book of the main results achieved within SIMBAD will foster further progress in this important emerging field.

References

1. Altschul, S.F., Gish, W., Miller, W., Meyers, E.W., Lipman, D.J.: Basic local alignment search tool. J. Mol. Biol. **215**, 403–410 (1990)

2. Balcan, M.F., Blum, A., Srebro, N.: A theory of learning with similarity functions. Mach. Learn. **72**(1–2), 89–112 (2008)
3. Biederman, I.: Recognition-by-components: a theory of human image understanding. Psychol. Rev. **94**, 115–147 (1987)
4. Bridgman, P.W.: The Logic of Modern Physics. MacMillan, New York (1927)
5. Bunke, H., Sanfeliu, A.: Syntactic and Structural Pattern Recognition: Theory and Applications. World Scientific, Singapore (1990)
6. Dubuisson, M.P., Jain, A.K.: Modified Hausdorff distance for object matching. In: Proc. Int. Conf. Pattern Recognition (ICPR), pp. 566–568 (1994)
7. Duda, R.O., Hart, P.E., Stork, D.G.: Pattern Classification. Wiley, New York (2000)
8. Easley, D., Kleinberg, J.: Networks, Crowds, and Markets. Cambridge University Press, Cambridge (2010)
9. Edelman, S.: Representation and Recognition in Vision. MIT Press, Cambridge (1999)
10. Goldstone, R.L., Son, J.Y.S.: In: Holyoak, K., Morrison, R. (eds.) The Cambridge Handbook of Thinking and Reasoning, pp. 13–36. Cambridge University Press, Cambridge (2005)
11. Jacobs, D.W., Weinshall, D., Gdalyahu, Y.: Classification with nonmetric distances: Image retrieval and class representation. IEEE Trans. Pattern Anal. Mach. Intell. **22**, 583–600 (2000)
12. Kleinberg, J.: Authoritative sources in a hyperlink environment. In: Proc. 9th ACMSIAM Symposium on Discrete Algorithms, pp. 668–677 (1998)
13. Lakoff, G.: Women, Fire, and Dangerous Things: What Categories Reveal About the Mind. University of Chicago Press, Chicago (1987)
14. Mayr, E.: The Growth of Biological Thought. Harvard University Press, Cambridge (1982)
15. Popper, K.R.: Conjectures and Refutations: the Growth of Scientific Knowledge. Routledge, London (1963)
16. Resnik, M.D.: Mathematics as a Science of Patterns. Clarendon, Oxford (1997)
17. Rorty, R.: A world without substances and essences. In: Philosophy and Social Hope, pp. 47–71. Penguin, London (1999)
18. Sen, P., Namata, G., Bilgic, M., Getoor, L., Gallagher, B., Eliassi-Rad, T.: Collective classification in network data. AI Mag. **29**(3), 93–106 (2008)
19. von Luxburg, U., Williamson, R.C., Guyon, I.: Clustering: Science or art? In: JMLR: Workshop and Conference Proceedings, vol. 27, pp. 65–79 (2012)
20. Watanabe, S.: Pattern Recognition: Human and Mechanical. Wiley, New York (1985)

Part I
Foundational Issues

Chapter 2
Non-Euclidean Dissimilarities: Causes, Embedding and Informativeness

Robert P.W. Duin, Elżbieta Pękalska, and Marco Loog

Abstract In many pattern recognition applications, object structure is essential for the discrimination purpose. In such cases, researchers often use recognition schemes based on template matching which lead to the design of non-Euclidean dissimilarity measures. A vector space derived from the embedding of the dissimilarities is desirable in order to use general classifiers. An isometric embedding of the symmetric non-Euclidean dissimilarities results in a pseudo-Euclidean space. More and better tools are available for the Euclidean spaces but they are not fully consistent with the given dissimilarities.

In this chapter, first a review is given of the various embedding procedures for the pairwise dissimilarity data. Next the causes are analyzed for the existence of non-Euclidean dissimilarity measures. Various ways are discussed in which the measures are converted into Euclidean ones. The purpose is to investigate whether the original non-Euclidean measures are informative or not. A positive conclusion is derived as examples can be constructed and found in real data for which the non-Euclidean characteristics of the data are essential for building good classifiers. (This chapter is based on previous publications by the authors, (Duin and Pękalska in Proc. SSPR & SPR 2010 (LNCS), pp. 324–333, 2010 and in CIARP (LNCS), pp. 1–24, 2011; Duin in ICEIS, pp. 15–28, 2010 and in ICPR, pp. 1–4, 2008; Duin et al. in SSPR/SPR, pp. 551–561, 2008; Pękalska and Duin in IEEE Trans. Syst. Man Cybern., Part C, Appl. Rev. 38(6):729–744, 2008) and contains text, figures, equations, and experimental results taken from these papers.)

R.P.W. Duin · M. Loog (✉)
Pattern Recognition Laboratory, Delft University of Technology, Delft, The Netherlands
e-mail: m.loog@tudelft.nl

R.P.W. Duin
e-mail: r.duin@ieee.org

E. Pękalska
Manchester, UK
e-mail: ela@elapekalska.com
url: http://www.elapekalska.com

M. Pelillo (ed.), *Similarity-Based Pattern Analysis and Recognition*,
Advances in Computer Vision and Pattern Recognition,
DOI 10.1007/978-1-4471-5628-4_2, © Springer-Verlag London 2013

2.1 Introduction

Automatic recognition systems work with objects such as images, videos, time signals, spectra, and so on. They are built in the process of learning from a set of object examples labeled with the desired pattern classes. Two main steps can be distinguished in this procedure:

Representation: Individual objects are characterized by a set of suitable mathematical descriptors such as vectors, strings of symbols or graphs. A good representation is the one in which objects can easily be related to each other in order to facilitate the next step.

Generalization/Discrimination: The representations of the object examples should enable the mathematical modeling of object classes or class discriminants such that a good class estimate can be found for new, unseen and, thereby, unlabeled objects using the same representation.

The most popular representations, next to strings and graphs, encodes objects as vectors in Euclidean vector spaces. Instead of single vectors, also sets of vectors may be considered for representing individual objects, as studied, e.g., in [32, 33, 46, 48]. For some applications, representations defined by strings of symbols and attributed graphs are preferred over vectors as they model the objects more accurately and offer more possibilities to include domain expert knowledge [6].

On the other hand, representations in Euclidean vector spaces are well suited for generalization. Many tools are available to build (learn) models and discriminant functions from sets of object examples (also called training sets) that may be used to classify new objects into the right class. Traditionally, the Euclidean vector space is defined by a set of features. These should ideally characterize the patterns well and be relevant for class differences at the same time. Such features have to be defined by experts exploiting their knowledge of the application.

The use of features has one important drawback. Features often represent the objects just partially because they encode their limited characteristics. Consequently, different objects may have the same representation, i.e., the same feature vector, when they differ by properties that are not expressed in the chosen feature set. This results in class overlap: in some areas of the feature space, objects of different classes are represented by the same feature vectors. Consequently, they cannot be distinguished any longer, which leads to an intrinsic classification error, usually called the Bayes error.

An alternative to the feature representation is the dissimilarity representation defined on direct pairwise object comparisons. If the entire objects are taken into account in the comparison, then only identical objects will have a dissimilarity zero (if the dissimilarity measure has the property of ‘identity of indiscernibles’). For such a representation class, overlap does not exist if the objects are unambiguously labeled, which means that there are no real world objects in the application that belong to multiple classes.

Some dissimilarity measures used in practice do not have the property that a zero dissimilarity can only arise for identical objects. An example is the single-linkage distance used in clustering: the dissimilarity between two clusters is defined

as the distance between the two most neighboring vectors. This distance measure corresponds to defining the smallest distance between the surfaces of two real world objects as the distance between the objects. A zero value, however, does not imply that the objects are identical; they are just touching.

Distance measures such as the above, and many others, cannot be perfectly embedded in a Euclidean space. This means that there is no set of vectors in a vector space of any dimensionality for which the Euclidean distances between the objects are identical to the given ones. In particular, it holds for non-metric distances, which are just an example from a large set of non-Euclidean distance measures. As we want to include non-metric distances (such as the single-linkage distance) we will use the more general term of dissimilarities instead of distances. They refer to possibly improper distance measures in the mathematical sense. We will still assume that dissimilarities are non-negative and that they have a monotonic relation with object differences: if two given objects are made more different, their dissimilarity increases.

Non-Euclidean symmetric dissimilarity data can be perfectly embedded into pseudo-Euclidean spaces. A proper embedding of non-Euclidean dissimilarities and the training of classifiers in the resulting space are, however, not straightforward. There are computational as well as fundamental problems to be solved. The question thereby arises whether the use of non-Euclidean dissimilarity measures is strictly necessary. Finding the causes of such measures, see Sect. 2.2, is a first step to answer this question. This will be more extensively discussed in Sect. 2.6. We will investigate whether such measures are really informative and whether it is possible to make Euclidean corrections or approximations by which no information is lost.

Two main vectorial representations of the dissimilarity data, the dissimilarity space and the pseudo-Euclidean embedded space, are presented in Sect. 2.3. Section 2.4 discusses classifiers which can be trained in such spaces. Transformations which make the dissimilarity data Euclidean are briefly presented in Sect. 2.5. Next, numerous examples of artificial and real dissimilarity data are collected in Sect. 2.7. Oftentimes, they illustrate that linear classifiers in the dissimilarity-derived vector spaces are much more advantageous than the traditional 1-NN rule. Finally, we summarize and discuss our findings in Sect. 2.8.

The issue of informativeness of the non-Euclidean measures is the main topic of this chapter. We will present artificial and real world examples for which the use of such measures is really informative. We will, however, also make clear that for any given classifier defined in a non-Euclidean space an equivalent classifier in a Euclidean space can be constructed. It is a challenge to do this such that the training of good classifiers in this Euclidean space is feasible. In addition, we will argue that the dissimilarity space as proposed by the authors [37, 55] is a Euclidean space that preserves all non-Euclidean information and enables the design of well performing classifiers.

2.2 Causes of Non-Euclidean Dissimilarities

In this section, we shortly explain why non-Euclidean dissimilarities frequently arise in the applications. This results from the analysis of a set of real world objects. Let D be an $N \times N$ dissimilarity matrix describing a set of pairwise dissimilarities between N objects. D is Euclidean if it can be perfectly embedded into a Euclidean space. This means that there exists a Euclidean vector space with N vectors for which all Euclidean distances are identical to the given ones.

There are N^2 free parameters if we want to position N vectors in an N-dimensional space. The dissimilarity matrix D has also N^2 values. D should be symmetric because the Euclidean distance is. Still, there might be no solution possible as the relation between vector coordinates and Euclidean distances is nonlinear. More on the embedding procedures is discussed in Sect. 2.3. At this moment, we need to remember that the matrix D is Euclidean only if the corresponding vector space exists.

First, it should be emphasized how common non-Euclidean measures are. An extensive overview of such measures is given in [55], but we have often encountered that this fact is not fully recognized. Most researchers wrongly assume that non-Euclidean distances are equivalent to non-metric ones. There are, however, many metric but non-Euclidean distances, such as the city-block or ℓ_1-norm.

Almost all probabilistic distance measures are non-Euclidean by nature. This implies that by dealing with object invariants, the dissimilarity matrix derived from the overlap between the probability density functions corresponding to the given objects is non-Euclidean. Also the Mahalanobis class distance as well as the related Fisher criterion is non-Euclidean. Consequently, many non-Euclidean distance measures are used in cluster analysis and in the analysis of spectra in chemometrics and hyperspectral image analysis as spectra can be considered as one-dimensional distributions.

Secondly, what is often overlooked is the following fact. One may compare pairs of real world objects by a (weighted) Euclidean distance, yet the complete set of N objects giving rise to an $N \times N$ dissimilarity matrix D is non-Euclidean. In short, this is caused by the fact that different parts or characteristics of objects are used per pair to define the object differences. Even if the dissimilarity is defined by the weighted sum of differences, as long as there is no single basis of reference for the comparison of *all pairs*, the resulting dissimilarity matrix D will be non-Euclidean. These types of measures often result from matching procedures which minimize the cost or path of transformation between two objects. Fundamental aspects of this important issue are extensively discussed in Sect. 2.2.2.3.

In shape recognition, various dissimilarity measures are based on the weighted edit distance, on variants of the Hausdorff distance or on nonlinear morphing. Usual parameters are optimized within an application w.r.t. the performance based on template matching and other nearest neighbor classifiers [14]. Almost all have non-Euclidean behavior and some are even non-metric [14].

In the design and optimization of the dissimilarity measures for template matching, their Euclidean behavior is not an issue. With the popularity of support vector

machines (SVMs), it has become important to design kernels (similarities) which fulfill the Mercer conditions [12]. This is equivalent to a possibility of an isometric Euclidean embedding of such a kernel (or dissimilarities). Next sections discuss reasons that give rise to violations of these conditions leading to non-Euclidean dissimilarities or indefinite kernels.

2.2.1 Non-intrinsic Non-Euclidean Dissimilarities

Below we identify some non-intrinsic causes that give rise to non-Euclidean dissimilarities. In such cases, it is not the dissimilarity measure itself, but the way it is computed or applied that causes the non-Euclidean behavior.

2.2.1.1 Numeric Inaccuracies

Non-Euclidean dissimilarities arise due to the numeric inaccuracies caused by the use of a finite word length. If the intrinsic dimensionality of the data is lower than the sample size, the embedding procedure that relies on an eigendecomposition of a certain matrix, see Sect. 2.3, may lead to numerous tiny negative eigenvalues. They should be zero in fact, but become nonzero due to numerical problems. It is thereby advisable to neglect dimensions (features) that correspond to very small positive and negative eigenvalues.

2.2.1.2 Overestimation of Large Distances

Complicated measures are used when dissimilarities are derived from raw data such as (objects in) images. They may define the distance between two objects as the length of the path that transforms one object into the other. Examples are the weighted edit distance [4] and deformable templates [31]. In the optimization procedure that minimizes the path length, the procedure may approximate the transformation costs from above. As a consequence, too large distances are found. Even if the objects are compared by a (weighted) Euclidean distance measure, the resulting set of dissimilarities in D will often become non-Euclidean or even non-metric.

2.2.1.3 Underestimation of Small Distances

The underestimation of small distances has the same result as the overestimation of large distances. It may happen when the pairwise comparison of objects is based on different properties for each pair, as it is the case, e.g., in studies on consumer preference data. Another example is the comparison of partially occluded objects in computer vision.

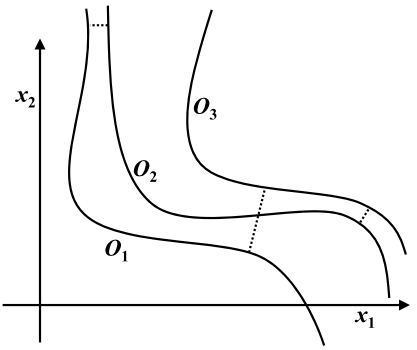

Fig. 2.1 Vector space with the invariant trajectories for three objects O_1, O_2 and O_3. If the chosen dissimilarity measure is defined as the minimum distance between these trajectories, triangle inequality can easily be violated, i.e., $d(O_1, O_2) + d(O_1, O_3) < d(O_1, O_3)$

2.2.2 Intrinsic Non-Euclidean Dissimilarities

The causes discussed in the above may be judged as accidental. They result either from computational or observational problems. If better computers and observations were available, they would disappear. Now, we will focuss on dissimilarity measures for which this will not happen. There are three possibilities.

2.2.2.1 Non-Euclidean Dissimilarities

As already indicated at the start of this section, arguments can be given from the application side to use another metric than the Euclidean one. An example is the l_1-distance between energy spectra as it is related to energy differences. Although the l_2-norm is very convenient for computational reasons and it is rotation invariant in a Euclidean space, other distance measures may naturally arise from the demands in applications, e.g., see [47].

2.2.2.2 Invariants

A fundamental reason behind non-Euclidean dissimilarities is related to the occurrence of invariants. Frequently, one is not interested in the dissimilarity between given objects A and B, but in the dissimilarity between their equivalence classes, i.e., sets of objects $A(\theta)$ and $B(\theta)$ in which θ controls an invariant. One may define the dissimilarity between the A and B as the minimum difference between the sets defined by all their invariants (see Fig. 2.1 for an illustration of this idea):

$$d^*(A, B) = \min_{\theta_A} \min_{\theta_B} \big(d\big(A(\theta_A), B(\theta_B)\big)\big). \tag{2.1}$$

This measure is non-metric: the triangle inequality may be violated as for different pairs of objects different values of θ are found minimizing (2.1).

2.2.2.3 Sets of Vectors

Complicated objects such as multi-region images may be represented by sets of vectors. Problems like this are investigated in the domain of Multi Instance Learning (MIL) [13], or Bag-of-Words (BoW) classification [52]. Distance measures between such sets have already been studied for a long time in cluster analysis. Many are non-Euclidean or even non-metric, such as the single linkage distance. This measure is defined as the distance between the two most neighboring points of the two clusters being compared. It is non-metric. It even holds that if $d(A, B) = 0$, then it does not follow that $A \equiv B$.

For the single linkage dissimilarity measure it can be understood why the dissimilarity space may be useful. Given a set of such dissimilarities between clouds of vectors, it can be concluded that two clouds are similar if the two sets of dissimilarities with all other clouds are about equal. If just their mutual dissimilarity is (close to) zero, they may still be very different.

The problem with the single linkage dissimilarity measure between two sets of vectors points to a more general problem in relating sets and even objects. In [33], an attempt has been made to define a proper Mercer kernel between two sets of vectors. Such sets are in that paper compared by the Hellinger distance derived from the Bhattacharyya's affinity between two pdfs $p_A(x)$ and $p_B(x)$ found for the two vector sets A and B:

$$d(A, B) = \left[\int \left(\sqrt{p_A(x)} - \sqrt{p_B(x)} \right)^2 \right]^{1/2}. \tag{2.2}$$

The authors state that by expressing $p(x)$ in any orthogonal basis of functions, the resulting kernel K is automatically positive semidefinite (psd). This is only correct, however, if all vector sets $A, B, \ldots$ to which the kernel is applied have the same basis. If different bases are derived in a pairwise comparison of sets, the kernel may become indefinite. This occurs if the two pdfs are estimated in a subspace defined by a PCA computed from the objects of the two classes A and B only.

This makes clear that indefinite relations may arise in any pairwise comparison of real world objects if every pair of objects if first represented in some joint space in which the dissimilarity is computed. These joint spaces may be different for different pairs! Consequently, the total set of dissimilarities will likely have a non-Euclidean behavior, even if each comparison relies on the Euclidean distance, as in (2.2).

The consequence of this observation is huge for pattern recognition applications. It implies that a representation defined by pairwise dissimilarities between objects can only be Euclidean if a common basis between all objects, including the future test objects, is found for the derivation of such dissimilarities. This is naturally, by definition, the case for feature vector representations, as the joint space for all objects is already defined by the chosen set of features. For the dissimilarity representation, however, which has the advantage of potentially using the entire objects,