Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Deliverable Responsible	EEMCS Delft University of Technology Mekelweg 4, 2628CD Delft The Netherlands http://ict.ewi.tudelft.nl/
Project web site	http://simbad-fp7.eu
EC project officer	Teresa De Martino
Document title	Foundations of (non)geometric similarities, final report
Deliverable n.	D3.4
Document type	Report
Dissemination level	Public
Contractual date of delivery	18
Project reference number	213250
Status & version	Final version
Work package	WP3
Deliverable responsible (SHORT NAME)	TUD
Contributing Partners (SHORT NAME)	TUD, ETH
Author(s)	R.P.W. Duin, J.M. Buhmann, E. Pękalska, V. Roth
Additional contributor(s)	M. Loog, W.R. Lee

Seventh Framework Programme
Information Communication Technologies
FET Open - Collaborative Project

Workpackage 3, Deliverable 3.4

Foundations of (non)geometric similarities

Robert P.W. Duin[1], Joachim M. Buhmann[2],
Elżbieta Pękalska[3], Volker Roth[4]

[1]Faculty of Electrical Engineering, Mathematics and Computer Sciences,
Delft University of Technology, The Netherlands
[2]Institute of Computational Science, ETH Zürich, Switzerland
[3]School of Computer Science, University of Manchester, United Kingdom
[4]Computer Science Department, University of Basel, Switzerland

April 2010

This is the final report of WP3. It summarizes the previous reports, hence it does not offer essential new information. See the bibliography for other reports and papers resulting from this workpackage.

1. Robert P.W. Duin, Wan-Jui Lee, Marco Loog and Elzbieta Pękalska, Study on (non)geometricity, SIMBAD Delivarable D3.1, April 2009.

2. Peter Schüffler, Sharon Wulff, Joachim M. Buhmann, Cheng Soon Ong, Volker Roth, Characterization of invariances, SIMBAD Delivarable D3.2, April 2009.

1 Introduction

The objective of this workpackage is to study both the causes of the lack of (geo)metricity in dissimilarity data and its effect on traditional machine learning algorithms.

For non-metric dissimilarity data, the triangle inequality is violated for some subsets of three data points. For metric data, a basic structure is available in the dataset. However, it is still possible that sets of more than three points would not fit into a Euclidean space. Consequently, they cannot be isometrically embedded in such a space. As a result, the assumptions are violated for many of the traditional procedures for analyzing and classifying data. For that reason the target of this workpackage is to study non-Euclidean data. We distinguish the following questions:

- Why is there non-Euclidean data?
- How can it be analyzed? What representation should be used?
- What are examples of datasets?

These are studied and presented in the three reports listed on page 2. The results are summarized in the next section. Consequences are discussed in the final section.

2 Results

2.1 The cause of non-Euclidean data

We identified two main causes for non-Euclidean behavior [1,4,5]: **non-intrinsic** ones and **intrinsic** ones.

Non-intrinsic causes are related to computational and observational problems. Examples are measurement noise, computational inaccuracies, algorithmic shortcuts in optimization procedures and missing data. The resulting dissimilarities may not be consistent with each other as they are measured/optimized only pairwise. In case there are no other effects, Euclidean representations can asymptotically be expected for increasing computational and observational resources.

Intrinsic causes are deliberately chosen for non-Euclidean distance measures like the l_1 norm used for comparing spectra, the edit distance and the Hausdorff distance used for comparing shapes and the single-linkage procedure used in cluster analysis (which is even non-metric). An important group of the set of intrinsic causes are the pairwise object comparisons that

define their own subspace or normalization [3,4,5]. Examples are objects represented by point sets in Hilbert space and rotation invariance of images solved by a pairwise optimization of the relative image angle. In such cases, a global representation for all objects might be derived by using all given pairwise relations leading to some non-Euclidean embedding. A perfect Euclidean embedding can only be found if the pairwise comparison was already made in a Euclidean space, which necessarily has to be constructed using all objects.

There is an important conclusion, not explicitly mentioned in one of the reports, but which is drawn here from the above. The use of training objects as well as test objects (i.e. the objects to be classified) for constructing the representation may be necessary in case the non-Euclidean dissimilarities are essential for a good classification. Transductive learning becomes then a consequence of using informative non-Euclidean object proximities [8].

An observation, not explicitly made before, but which is evident from all examples is the following. A basic reason for experts to design non-Euclidean dissimilarity measures is when it is essential to incorporate object structure into a measure. The bridge over the gap between structural and statistical pattern recognition that is offered by the dissimilarity representation demands the handling of non-Euclidean data as a toll [4].

2.2 The representation of non-Euclidean data

We can distinguish three approaches to construct a vector space given a full set of dissimilarities.

- **Embedding** into a pseudo-Euclidean space. This is possible without any error, so there is no loss of information. A problem is however that the inner product definition and the distance definition demand different classifiers in this space than the traditional ones. The SVM suffers from non-Mercer or indefinite kernels, density estimation is not well defined and objects can have negative square distances to other objects. Some classifiers can be defined, e.g. the nearest neighbor, nearest mean and Parzen [8]. For other classifiers, like QDA an Fisher, indefinite kernelized versions exist [6,7].

- **Euclidean corrections**, either based on the pseudo-Euclidean embedding, or directly on the given dissimilarity matrix [8,10,11] are available as well. Examples are neglecting the specific distance measure defined for the pseudo-Euclidean space or using a Euclidean subspace of this space. An interesting alternative is the transformation studied in report

D3.2 [2]: enlarging the off-diagonal dissimilarities by some constant before embedding. A perfect Euclidean embedding is possible for sufficiently large values of this constant. As this transformation is topology preserving (neighborhood relations are maintained, certain clusterings are invariant), it shows that a continuous (nonlinear!) transformation between classifiers in the pseudo-Euclidean space and classifiers in the Euclidean space exists that assign objects to classes in the same way.

- **The dissimilarity space**. This approach postulates a Euclidean space using the vector of dissimilarities from a given object to the so-called representation set (e.g. the training set) as 'features' [1,3,4]. This can always be done without any computational overhead, except when dimension reduction is desired. It makes no difference whether the original set of dissimilarities has a Euclidean behavior or not. It is even not needed that the pairwise relations are symmetric. The price that has to be paid for this flexible solution is that the distances in the dissimilarity space are not identical to, and can be very different from the original dissimilarities. This may be bad for some applications, but it can also be advantageous, e.g. in the above mentioned situation when the pairwise object comparisons made for the computation of the original dissimilarities do not take into account the context of the total set of objects. The dissimilarity space creates such a context by relating all objects to each other.

2.3 Examples of datasets

The datasets collected in the project are described in the SIMBAD report 2009_9 [8]. We collected 64 dissimilarity matrices of which 44 have been generated for the same problem (shapes of chicken pieces). This is a public domain data. From the remaining 20 matrices, six matrices are artificially generated by us, three matrices have been constructed within SIMBAD by using the COIL dataset, five matrices are the result of in-house applications and six matrices are public domain datasets taken from the internet. All have non-Euclidean behavior.

We developed software for a systematic analysis of the datasets and compared the performance of some classifiers in various spaces as discussed above. In general, dissimilarity spaces do somewhat better than pseudo-Euclidean embedded spaces and spaces derived by Euclidean corrections.

The main finding however is that for some datasets a direct removal of the non-Euclidean characteristics of the data (neglecting eigenvectors with negative eigenvalues in the pseudo-Euclidean embedding) is counterproductive

for some classifiers. So, the non-Euclidean behaior is informative for these datasets and these classifiers. This does not necessarily imply that classifiers used in the pseudo-Euclidean space or based on indefinite kernels are only beneficial. Euclidean corrections may bring the relevant information into the domain of Euclidean classifiers, as well.

Extreme examples are artificially constructed by the Balls3D and Balls50D datasets [5,8]. For these problems, all information for the separation between the classes is located in the negative part of the pseudo-Euclidean space. By isolating this area and treating it as a Euclidean space, the classification problems can be entirely solved.

One of the datasets that has become available in the SIMBAD project (WP7) is the Verona set of 182 dissimilarity matrices derived from the MRI brain scans of 124 subjects [9]. They are constructed using 13 different dissimilarity measures applied to ROIs related to 14 different regions in the brain. Majority of the measures is non-Euclidean. The data is analyzed with the same procedures as mentioned above. In addition, they are combined into a single dissimilarity matrix which showed a significantly better performance than any of the individual constituting matrices.

3 Conclusions

Causes for the non-Euclideaness of dissimilarity data have been identified. Some are non-intrinsic causes and corrections may improve classification results in such cases. For intrinsic causes, the non-Euclidean behavior is likely to be informative and Euclidean corrections or the construction of the Euclidean dissimilarity space may make the standard statistical classifiers applicable to these problems. We identified structural pattern recognition as a significant application domain of non-Euclidean dissimilarity measures, offering a bridge to the tools of statistical pattern recognition [16,17]. Examples are the many applications using dissimilarity measures between spectra or histograms like a shape distance or the earth mover distance [22,23,24,26,27,28].

4 Bibliography

In this section we will shortly discuss all reports and papers resulting from WP3 of the SIMBAD projects. Some of them resulted from collaboration with WP4 and WP7. Others are the result of collaboration with colleagues outside SIMBAD but in which the research supported by SIMBAD played a significant role.

These are the two deliverables of the first year of WP3. They describe the causes the give rise to indefinitie and non-Euclidean data and discuss extensively the circumstances under which such data can be transformed into an Euclidean space without affecting its cluster structure.

1. R.P.W. Duin, W.J. Lee, M. Loog and E. Pękalska, Study on (non)geometricity, SIMBAD Deliverable D3.1, April 2009.

2. P. Schüffler, S. Wulff, J.M. Buhmann, C.S. Ong, V. Roth, Characterization of invariances, SIMBAD Deliverable D3.2, April 2009.

The findings and results of Deliverable D3.1 are crystallized in the following papers and a tutorial. This tutorial has been presented at SCIA 2009 in Oslo, Norway and at two Colombian universities in Manizales and Cali in October 2009.

3. R.P.W. Duin and E. Pękalska, The dissimilarity representation for pattern recognition, a tutorial, 2009 (Simbad Technical Report n. 2009_x).

4. R.P.W. Duin, Pattern Recognition as a Human Centered non-Euclidean Problem, (invited) in: Proc. ICEIS 2010, Funchal, Madeira, Portugal, June 2010 (Simbad Technical Report n. 2009_x).

5. R.P.W. Duin and E. Pękalska, Non-Euclidean Dissimilarities: Causes and Informativeness, Proc. SSSPR, 2010 (Simbad Technical Report n. 2009_x).

The following two papers cannot be based as output of the SIMBAD project as none of the authors is really supported by it. Elzbieta Pękalska however participates in the project and the papers are of significant importance as they deal with the use of indefinite kernels for implicitly building classifiers in the pseudo-Euclidean space.

6. B. Haasdonk and E. Pękalska: Indefinite Kernel Fisher Discriminant. ICPR 2008.

7. E. Pękalska and B. Haasdonk, Kernel discriminant analysis with positive definite and indefinite kernels. IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 10171032.

The below two extensive SIMBAD reports are the results of a systematic analysis of 64 general datasets and 182 MRI datasets (in cooperation with WP7).

8. R.P.W. Duin and E. Pękalska, Datasets and tools for dissimilarity analysis in pattern recognition, Simbad Technical Report n. 2009_9, 174 pages, November 2009

9. R.P.W. Duin, Aydin Ulas, E. Pękalska, V. Murino and P. Brambilla, The Verona MRI brain images, a first analysis of the dissimilarities, Simbad Technical Report n. 2010_x, 460 pages, April 2010.

Some papers on Euclidean corrections and other transformations of dissimilarity matrices.

10. R.P.W. Duin, E. Pękalska, A. Harol, W.J. Lee, and H. Bunke, On Euclidean corrections for non-Euclidean dissimilarities, in: N. da Vitoria Lobo, T. Kasparis, F. Roli, J.T. Kwok, M. Georgiopoulos, G.C. Anagnostopoulos, M. Loog (eds.), Structural, Syntactic, and Statistical Pattern Recognition, Proc. SSSPR2008 (Orlando, Plorida, USA, 4-6 Dec 2008), Lecture Notes in Computer Science, vol. 5342, Springer Verlag, Berlin, 2008, 551-561.

11. R.P.W. Duin and E. Pękalska, On refining dissimilarity matrices for an improved NN learning, Proc. of the 19th Int. Conf. on Pattern Recognition (ICPR2008, Tampa, USA, December 2008), IEEE Press, 2008.

In several papers combining is studied of sets of dissimilarity representations for the same objects resulting from different dissimilarity measures.

12. S.W. Kim and R.P.W. Duin, A Combine-Correct-Combine Scheme for Optimizing Dissimilarity-Based Classifiers, in: E. Bayro-Corrochano, J.O. Eklundh (eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (Proc. CIARP 2009, Guadalajara, Jalisco, Mexico, Nov. 15-18, 2009), Lecture Notes in Computer Science, vol. 5856, Springer, Berlin, 2009, 425-432.

13. S.W. Kim and R.P.W. Duin, An emperical comparison of kernel-based and dissimilarity-based feature spaces, Proc. SSSPR, 2010.

14. A. Ibba, R.P.W. Duin, and W.J. Lee, A study on combining sets of differently measured dissimilarities, ICPR 2010, 2010 (Simbad Technical Report n. 2009_x).

15. W.J. Lee, R.P.W. Duin, M. Loog, and A. Ibba, An Experimental Study On Combining Euclidean Distances, Proc. The 2nd International Workshop on Cognitive Information Processing (14-15-16 June, 2010 Elba Island, Tuscany - Italy), 2010 (Simbad Technical Report n. 2009_x).

The next papers (partly based on WP7 research) study ways to use the dissimilarities for embedding sets of (attributed) graphs. The resulting dissimilarity matrices are almost always non-Euclidean.

16. W.J. Lee and R.P.W. Duin, A Labelled Graph Based Multiple Classifier System, in: J.A. Benediktsson, J. Kittler, F. Roli (eds.), Multiple Classifier Systems (Proc. 8th Int. Workshop, MCS 2009, Reykjavik, Iceland, June 10-12, 2009), Lecture Notes in Computer Science, vol. 5519, Springer, Berlin, 2009, 201-210.

17. W.J. Lee, R.P.W. Duin, and H. Bunke, Selecting Structural Base Classifiers for Graph-based Multiple Classifier Systems, in: N. El Gayar, J. Kittler, F. Roli (eds.), Multiple Classifier Systems (Proc. 9th Int. Workshop, MCS 2010, Cairo, Egypt), Lecture Notes in Computer Science, vol. 5997, Springer, Berlin, 2010, 155-164 (Simbad Technical Report n. 2009_x).

Dissimilarity representations are applied to the areas of missing data, prototype selection and multi-instance learning.

18. M. Millan-Giraldo, R.P.W. Duin, and J.S. Sanchez, Dissimilarity-based Classification of Data with Missing Attributes, Proc. The 2nd International Workshop on Cognitive Information Processing (14-15-16 June, 2010 Elba Island, Tuscany - Italy), 2010.

19. Y. Plasencia-Calana, E. García-Reyeso, M. Orozco-Alzate, and R.P.W. Duin, Prototype selection for dissimilarity representation by a genetic algorithm, ICPR 2010, 2010.

20. L. Sørensen, R.P.W. Duin, M. de Bruijne, W.J. Lee, D.M.J. Tax, and M. Loog, Dissimilarity-based Multiple Instance Learning, Proc. SSSPR, 2010.

The usage of dissimilarities for various applications has been studied: thermal image recognition, chemometrics, MRI based dementia diagnosis, MRI based detection of schizophrenia, volcanic seismics, and the diagnosis of lung diseases.

21. Y. Plasencia, E. García-Reyes, R. P. W. Duin, H. Mendez-Vazquez, C. San-Martin, and C. Soto, A Study on Representations for Face Recognition from Thermal Images, in: E. Bayro-Corrochano, J.O. Eklundh (eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (Proc. CIARP 2009, Guadalajara, Jalisco, Mexico, Nov. 15-18, 2009), Lecture Notes in Computer Science, vol. 5856, Springer, Berlin, 2009, 185-192.

22. D. Porro Muñoz, R.P.W. Duin, I. Talavera, and N. Hernández, The Representation of Chemical Spectral Data for Classification, in: E. Bayro-Corrochano, J.O. Eklundh (eds.), Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications (Proc. CIARP 2009, Guadalajara, Jalisco, Mexico, Nov. 15-18, 2009), Lecture Notes in Computer Science, vol. 5856, Springer, Berlin, 2009, v513-520.

23. D. Porro Muñoz, R.P.W. Duin, M. Orozco-Alzate, I. Talavera-Bustamante, and J.M. Londoño-Bonilla, The Dissimilarity Representation as a Tool for Three-way Data Classification: a 2D Measure, Proc. SSSPR, 2010.

24. D. Porro-Muñoz, I. Talavera, R.P.W. Duin, N. Hernández, and M. Orozco-Alzate, Dissimilarity Representation on Functional Spectral Data for Classification, Journal of Chemometrics, 2010, submitted.

25. S. Klein, M. Loog, F. van der Lijn, T. den Heijer, A. Hammers, M. de Bruijne, A. van der Lugt, R.P.W. Duin, M.M.B. Breteler, and W.J. Niessen, Early diagnosis of dementia based on intersubject whole-brain dissimilarities, Proc. of IEEE International Symposium on Biomedical Imaging: Macro to Nano (Rotterdam, 14-17 April 2010), 2010.

26. A. Ulas, R.P.W. Duin, U. Castellani, M. Loog, M. Bicego, V. Murino, M. Bellani, S. Cerruti, M. Tansella, and P. Brambilla, Dissimilarity-based Detection of Schizophrenia, Proc. ICPR 2010 workshop on Pattern Recognition Challenges in FMRI Neuroimaging, 2010, submitted.

27. D. Porro-Muñoz, R.P.W. Duin, M. Orozco-Alzate, I. Talavera, and J.M. Londoño- Bonilla, Classifying Three-way Seismic Volcanic Data by Dissimilarity Representation, ICPR 2010, 2010.

28. L. Sørensen, M. de Bruijne, R.P.W. Duin, M. Loog, P. Lo, and A. Dirksen, Image Dissimilarity-Based Quantification of Pathology, MICCAI, 2010.