Final work package report

Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Deliverable Responsible	UNIVERSITY OF YORK Department of Computer Science Heslington Hall, United Kingdom, YO10 5DD
Project web site	http://simbad-fp7.eu
EC project officer	Teresa De Martino
Document title	Imposing geometricity on non-geometric similarities (embedding)
Deliverable	D4.4
Document type	Final Report
Dissemination level	Public
Contractual date of delivery	M 30
Project reference number	213250
Status & version	Definitive version
Work package, deliverable responsible	WP 4, UNIYORK
Contributing Partners	DELFT, ETH, UNIVE
Author(s)	Edwin Hancock, Richard Wilson, Joachim Buhmann, Bob Duin and Andrea Torsello
Additional contributor(s)	-

Final Report: WP4 - Imposing geometricity on non-geometric similarities (embedding)

Edwin Hancock, Richard Wilson, Joachim Buhmann, Bob Duin and Andre Torsello

September 19, 2011

1 Introduction

The technical annex defined the goals of WP4 as follows:

"The basic assumption underlying the work within this workpackage is that similarity data is given, possibly in the form of a weighted graph, and we aim at developing algorithms for transforming them into instance-specific vectorial representations (embedding) that are suitable for traditional geometric learning algorithms."

The workpackage was divided into three subpackages, each with its own deliverable

- WP4.1: Spectral and geometric manifold embedding
- WP4.2: Structure-preserving embeddings
- WP4.3: Graph regularisations

The overall aim in task WP4.1 was to develop spectral methods for embedding weighted graphs in a geometrically meaningful way, and using the resulting embeddings to construct generative models for graph structure. In particular, the aim was to develop spectral methods for embedding with guaranteed curvature properties.

Task WP4.2 addresses a second category of embedding methods with a fundamentally different focus. Instead of approximating the original (dis)similarities by Euclidean distances, these approaches try to preserve the underlying group structure of the data. Here our aim was to use the analysis of the model invariances developed in WP3 to devise model-specific embedding procedures that preserve the structure underlying specific machine learning tasks rather than the actual dissimilarities. Such distortion-free embeddings then allow us to apply the whole arsenal of data preprocessing methods that have been developed for vectorial data over the last decades.

Finally, WP4.3. had twofold aims. The first of these was to study the relations between our generative models and the generative kernels studied in WP2.1. Besides capturing the modes of variation present in sets of graphs it is also important to be able to smooth or regularize/denoise them. To this end, we planned to study the effects of the heat or diffusion equation in the tangent space produced by the exponential map. Finally, we aimed to undertake an analysis of spectral-geometric corrections of the lack of metricity of the data.

2 Progress

The main scientific achievements made in the different workpackages are described in the relevant deliverable. However the main contributions (with reference to the published results from the project) are as follows

2.1 WP4.1: Spectral Embedding

- Use of the Ihara zeta function as a means of embedding graphs [2], weighted graphs and hypergraphs [1] in a vector space. Analysis of this representation and its links to classical and quantum walks [3]. Development of an efficient means to evaluate the Ihara zeta function through Bell polynomial recursion. Undertaken partly in collaboration with Venice [4].
- Constant curvature spherical embeddings for non-Euclidean data [9]. Development of classifiers based on tangent plane reprojection of data, using Lie group (Exp and Log map) representation of embedded data. Undertaken jointly with Delft [9, 16].
- Application of Ricci flows to the embedded data, with the aim of reducing the mass of negative eignevalues through manifold flattening [10, 15, 20, 23].
- Study of the complexity of similarity representations and their embeddings via the use of information theoretic measures (Shannon enttropy, Non Neumann entropy, thermodynamic depth complexity [5].

2.2 WP4.2: Structure preserving embeddings

In the first stage of this workpackage a novel extension of the Wishart-Derichlet cluster process was developed to deal with disimilarity data.

In the final stage of this subproject 4.2 Zurich have investigated the unsolved problem of validating spectral clustering principles. This task consists of (i) ranking different clustering objectives according to their informativeness, and (ii) determining the most appropriate number of clusters. A general validation principle for clustering cost functions has been derived on the basis of information theory [28]. Informative and reliable clustering models result in maximal approximation capacity. This deliverable demonstrates the principle of maximum approximation capacity for two spectral clustering models: correlation clustering and pairwise clustering. These models have been introduced in the previous deliverable and they are used to analyze the correlations of temporal gene expression profiles. Experimental results demonstrate that pairwise clustering yields significantly more informative cluster structures than correlation clustering. The obtained results are consistent with BIC, but the validation method is more generally applicable than BIC regularized clustering.

The validation principle can be viewed as a monitoring instrument which enables us to measure the information content of objective functions. Clustering objectives with additional properties like "being embeddable in Euclidean spaces" without distortions of the clusterings might reduce the informativity of clusterings. Our investigation produced the surprising result that pairwise clustering which can be reduced to k-means clustering in a sufficiently high-dimensional space has a higher informatiuon content than correlation clustering without this additional embedding property. Being embeddable seems to also preserve information. These results are summarized in a technical report.

2.3 WP4.3: Regularisation and Kernels

- Regularisation of the Ricci flow embedding using both heat-kernel and p-Laplacian methods. Both methods appear to reduce the mass of negative eignenvalues in the updated disimularity matrix [6, 7, 8]. Development of tangent-space classifiers in conjunction with the regularisation [20].

- Exploration of alternative spectral representations and smoothing operators based on the edge-based Laplacian, including the heat and wave equations [6, 7, 8, 14].

- Developing new computationally efficient measures of graph, namley the thermodynamic depth complexity [5], and the Von-Neumann entropy [17, 18]. The thermodynamic depth complexity, associates with heat flow on a graph the concept of node entropy history, and this together with the notion of phase-change can be used to assign a complexity measure to a graph. The second approach has been to develop a simplified computation of the Von Neumann entropy from quantum systems. This is the Shannon entropy associated with the normalised Laplacian spectrum. By making a quadratic approximation to the Shannon entropy, we have shown how to approximate the Von Neumann entropy using simple node degree statistics in quadratic time.

- Use of the above complexities to design graph kernels using Shannon-Jensen framework of WP2 from Lisbon [22].

- Use of the above complexity measures as measures of model order complexity in the learning of generative models of graph structure [24]. Links with the hypothesis testing framework from Zurich.

- We have also addressed the problem of learning a generative model of both the structure and attributes observed in a set of graphs [26]. This has been approached so as to better capture underlying structure of the representation and provide a better view of the real similarities of the graphs. In this way we dispense with the isotropy assumption underlying many approaches to graph-matching. We present a naive node-observation model, where we make the important assumption that the observation of each node and each edge is independent of the remainder. We then propose an EM-like approach to learn a mixture of these models and a Minimum Message Length criterion for component selection. Moreover, in order to avoid the bias that could arise with a single estimation of the node correspondences, we opt to estimate the sampling probability over all the possible matches. The resulting classifier greatly outperforms the nearest neightbour rule on all datasets tested regardless of the graph-similarity measure adopted.

3 Advances to the State-of-the-art

Taken together this work can be regarded as advancing the state of the art in the following wats

- Developing techniques for embedding non-Euclidean data onto constant curvature manifolds of appropriate signed curvature (either elliptic or hyperbolic). Using this type of embedding, non-Euclidean data can either be clustered on the surface of the manifold, or projected into a tangent space where standard classifiers can be deployed.

- Providing new methods for embedding a variety of structural representations (graphs, weighted graphs and hypergraphs) based on deeper measures of structure (prime cycle frequencies). Establishing the links between this represntation and classical/quantum walks on graphs.

- Providing methods based on structure to embed both similarity and graph data into vector spaces. We have explored both path-based and cycle-based methods. Here the Ihara zeta function has opened up new possibilities for analysing graph data that spans both the spectral and structural characterisation. We have also used a novel extension of the Wishart-Derichlet cluster process to deal with disimilarity data.

- Providing new information theoretic methods for designing graph kernels and learning generative models of both structure and attributes based on information theoretic measures of structural complexity. This work on learning with graphs extends the construction of generative models from the tree to the graph domain. The work on kernels, although at an early stage provides a promissing approach to constructing a new family of graph kernels.

- We have explored a variety of graph regularisation strategies. For embedded data, methods based on Ricci flow smoothing have been shown to reduce the non-Euclidean effects associated with negative eigenvalues of the dissimilarity matrix. We have also shown that graph-diffusion processes to be readily extendable to more general pairwise geometric constraints [25], effectively generalizing the similarity- based approach to any set of pairwise spatial constraints. Finally, we have explored alternative regularisation methods based on the edge-based Laplacian rather than the node-based Laplacian.

- Most spectral embeddings in the literature are only suitable for graphs with no labels/attributes, and it is one of the most significant drawback of spectral embeddings. To overcome this limitation of spectral embeddings, we proposed a solution [27] for embedding labeled graphs with spectral methods by transforming a labeled graph into a new set of unlabeled graphs and preserving all the linkages at the same time. By embedding the label information into edges, we can further ignore the labels. By assigning weights to the edges according to the labels of their linked nodes, the strengths of the connections are altered, but the topology of the graph as a whole is preserved. Then, any spectral embedding method can be applied on these unlabeled graphs which are transformed from labeled graphs.

Taken together, we believe the impact of these methods have the following impact on the scientific community.

First, from a pragmatic standpoint we have provided a variety of techniques that can be used to appraise similarity data, assess the potential cause of non-Euclidean artefacts and render the data amenable to analysis. Moreover, we have provided a set of techniques that can be used to embed this data into vector spaces and onto appropriately chosen manifolds, where non-Euclidean effects can be controlled and the analysis of data performed. Finally, we have provided techniques for regularising disimilarity data and for reducing the contributions from negative eigenvalues.

Second, we have provided new theoretical insights as to how the embedding of disimilarity data should be conducted. Here we have taken both spectral and structural approaches. In both areas we have developed novel methodologies, which open up new research directions.

4 Collaborations

4.1 Internal

The following collaborations took place during the workpackage

- York and Delft collaborated on the use of spherical embeddings for analysing non-Euclidean similarity data. This resulted in a joint paper to CVPR 2010. This work was initiated by a visit of Duin and Pekalska to York in July 2008. It was further developed at the Castelbrando workshop, and has lead to a journal submission that is currently being finalised.

- Samuel Rota-Bula visited York in June 2011 to work on research related to hypergraphs and the Ihara zeta function. This resulted in a joint York/Venice paper on efficiently evaluating the Ihara zeta function via Bell polynomial recursion, and this is accepted to appear in Linear Algebra and Applications.

- Richard Wilson and Edwin Hancock visited Lisbon in July 2009 to discus information theoretic and generative kernels with Mario Figueiredo, and graph-based clustering with Ana Fred. These ideas were further persued at the Castelbrando workshop, and Lin Han and Lu Bai from York have recently reported early results on the use of entropy measures to define Jensen-Shannon kernels.

- Lin Han from York visited Zurich in September 2011 to discus links between her information theoretic approach to generative models and kernels for graphs, and the Zurich hypothesis testing framework. This has opened up the possibility of applying this work to the Zurich datasets, and this work will be reported in her PhD thesis that will appear in early 2012.

4.2 External

- The work on complexity measures based on polytopal expansions and thermodynamic depth was done in conjunction with scientists at the University of Alicante (Francesco Esclolano and Miguel-Algelo Lozano). This work has recently lead to a CVPR 2011 paper on using entropy measures for embedded graph matching via alignment [19].

References

[1] Peng Ren, T. Aleksic, R.C. Wilson and E.R. Hancock “A Polynomial Characterization of Hypergraphs Using The Ihara Zeta Function”, *Pattern Recognition*, **44**, pp. 1941-1957, 2011.

[2] Peng Ren, R.C. Wilson and E.R. Hancock, “Graph Characterization via Ihara Coefficients”, *IEEE Transactions of Neural Networks*, **22**, 233-245, 2011.

[3] Pen Ren, T. Aleksic, R.C. Wilson and E.R. Hancock, “Ihara Zeta Functions, Quantum Walks and Cospectrality in Strongly Regular Graphs”, *Quantum Information Processing*, **10**, pp.405–417, 2011.

[4] S Rota-Bulo, E.R. Hancock, F. Aziz and M. Pelillo, “Efficient Computation of Ihara co-efficients using the Bell Polynomial Recursion”, Linear Algebra and Applications, to appear.

[5] F. Escolano, M-A. Lozano and E.R. Hancock “The Heat Diffusion - Thermodynamic Depth Complexity of Networks”, Phys. Rev. E, submitted.

[6] H. ElGhawalby and E.R. Hancock, “Graph characteristic from the Gauss Bonnet Theorem”, SSPR 2008, Lecture Notes in Computer Science, **5342**, pp. 207–216, 2008.

[7] H. ElGhawalby and E.R. Hancock, “Characterizing graphs using spherical triangles”, IbPRIA 2009, Lecture Notes in Computer Science, **5524**, pp. 465–472, 2009.

[8] H. ElGhawalby and E.R. Hancock, “Graph Regularisation using Gaussian Curvature”, GbR 2009, Lecture Notes in Computer Science, **5534**, pp. 233–242, 2009.

[9] R. C. Wilson, E.R. Hancock, E. Pekalska and R. Duin, “Spherical Embeddings for non-Euclidean Dissimilarities”, CVPR pp. 1903–1910, 2010.

[10] W. Xu, R.C. Wilson and E.R. Hancock, “Rectifying Non-Euclidean Similarity Data using Ricci Flow Embedding”, ICPR, pp. 3324-3327, 2010.

[11] L. Han, R.C. Wilson and E.R. Hancock, “A Supergraph-based Generative Model”, ICPR, pp. 1566-1569, 2010.

[12] F. Escolano, M.A, Lozano and E.R. Hancock, “Heat Flow-Thermodynamic Depth Complexity in Networks”, ICPR, pp. 1578-1581, 2010.

[13] P. Ren, T. Aleksic, R.C. Wilson and E.R. Hancock “Ihara Coefficients: A Flexible Tool for Higher Order Learning”, S+SSPR, LNCS **6218**, 670-679, 2010.

[14] H ElGhawalby and E.R. Hancock, “Graph Embedding using an Edge-based Wave Kernel, S+SSPR, LNCS **6218**, pp. 60-69, 2010.

[15] W. Xu, R.C. Wilson and E.R. Hancock, “Regularisng the Ricci Flow Embedding”, S+SSPR, LNCS **6218**, pp. 579–588, 2010.

[16] R.C. Wilson and E.R. Hancock, “Spherical Embedding and Classification”, S+SSPR, LNCS **6218**, pp. 589–599, 2010.

[17] L. Han, E.R. Hancock and R.C. Wilson, “Entropy versus Heterogeneity for Graphs”, GbRPR 2011, LNCS 6658, pp. 32-41, 2011.

[18] L. Han, E.R. Hancock and R.C. Wilson, “Learning Generative Graph Prototypes Using Simplified von Neumann Entropy”, GbRPR 2011, LNCS 6658, pp. 42-51, 2011.

[19] F. Escolano, E.R. Hancock and M.A. Lozano, “Graph Matching through Entropic Manifold Alignment”, CVPR, pp. 2417-2424, 2011.

[20] W. Xu, E.R. Hancock and R.C. Wilson, “Rectifying Non-euclidean Similarity Data through Tangent Space Reprojection”, IbPRIA 2011, LNCS 6669, pp. 379-386, 2011.

[21] L. Han, E.R. Hancock and R.C. Wilson, “Characterizing Graphs Using Approximate von Neumann Entropy”, IbPRIA 2011, LNCS 6669, pp. 484-491, 2011.

[22] L. Bai and E.R. Hancock “Graph Clustering using the Jensen-Shannon Kernel”, CAIP 2011, LNCS 6854, pp. 394–401, 2011.

[23] W. Xu, R.C. Wilson and E.R. Hancock, “Determining the Cause of Negative Dissimilarity Eigenvalues”, CAIP 2011, LNCS 6854, pp. 589–597, 2011.

[24] L. Han, R.C. Wilson and E.R. Hancock, “An Information Theoretic Approach to Learning Generative Graph Prototypes”, SIMBAD 2011, LNCS 7005, pp. 136-151, 2011.

[25] A. Torsello, E. Rodo, A. Albarelli, ”Multiview Registration via Graph Diffusion of Dual Quaternions.” In IEEE International Conference on Computer Vision and Pattern Recognigion, IEEE Computer Society, 2011

[26] A. Torsello and L. Rossi, ”Supervised Learning of Graph Structure.” To appear in Proc. SIMBAD2011, LNCS, Spriger.

[27] R. P. W. Duin W.-J. Lee and H. Bunke. Selecting structural base classifiers for graph-based multiple classifier systems. In 9th International Workshop on Multiple Classifier Systems, Lecture Notes on Computer Science, number 5997, pages 155–164, 2010.

[28] M. H.Chehreghani, A.G. Busetto and J.M. Buhmann, “Information Theoretic Model Validation for Spectral Clustering”, (in preparation), 2011.