Project acronym	SIMBAD
Project full title	Beyond Features: Similarity-Based Pattern Analysis and Recognition
Deliverable Responsible	Dep. de Engenharia Electrotécnica e ComputadoresInstituto Superior Técnico1049-001 Lisboa (Portugal) http://www.ist.utl.pt
Project web site	http://simbad-fp7.eu
EC project officer	Teresa De Martino
Document title	Deriving Similarities for non-Vectorial Data
Deliverable n.	D2.4
Document type	Report
Dissemination level	Public
Contractual date of delivery	M 42
Project reference number	213250
Status & version	Definitive version
Work package	WP 2
Deliverable responsible	IST
Contributing Partners	UNIVE, UNIVR, DELFT
Author(s)	Mário Figueiredo, Ana Fred
Additional contributor(s)	Vittorio Murino, Manuele Bicego

Deliverable D2.4

Work Package WP2: "Deriving similarities for non-vectorial data"

Final Report

The main goal of WP2 ("Deriving similarities for non-vectorial data") was to develop kernels and more general similarity measures for non-vectorial data. In this third and final report, we describe the main achievements in each of the tasks in which WP2 is divided (WP2.1 "Generative Kernels", WP2.2 "Compression Kernels", and WP2.3 "Learning and Combining Similarities"), and point to the resulting publications.

1. Task WP2.1: Generative Kernels

Generative kernels and information-theoretic kernels (the topic of WP2.2) are closely related. Both are based on the assumption that the objects of interest where generated by a probabilistic mechanism - a source, in information theoretic terms - and then proceed by defining (dis)similarity measures or kernels between models of these probabilistic sources. In fact, information-theoretic kernels can be considered as generative kernels, although the information-theoretic perspective tends to be more agnostic; i.e., it does not assume that the adopted model reflects the truth. In this section, we will describe the work on the topic traditionally called "generative kernels", and will postpone the work on information theoretic kernels to the section devoted to task WP.2.2.

The lead contributor to the WP2.1 task was the UNIVR partner, but in the second and third year with significant integration with the information-theoretic kernels developed in the first year in the context of WP2.2; the results of this combination of generative approaches with information-theoretic kernels will be described in the section devoted to WP2.2.

The research work on generative kernels has mainly focused on three directions:

(i) learning the generative model underlying a generative kernel (for example, for using in Fisher kernels);

(ii) the development and investigation of new score spaces based on the notion of *free energy* of a generative model;

(iii) the study on the impact of the normalization step of the score space underlying a typical generative kernel.

In the first work direction, a study of the best procedure to build generative models for Fisher kernels was carried out. In particular, we started from the observation that, in the context of generative kernels, most of the research efforts have been devoted to the discriminative step, namely to the definition of a proper score space or of a kernel to be used with a SVM.

Actually, in a typical scenario, a single generative model describing the whole problem is employed. Alternatively, approaches using two models (one for the positive class and another for the rest) or one model per class have shown to increase the performances. We went one step ahead in this direction, allowing a generative framework to freely discover the natural structures or groups in the training set. This was achieved with a preliminary step of clustering, during which a large number of small hidden natural groups is extracted from the data, disregarding class label information. Subsequently, a single and simple generative model is trained for each group (as the groups tend to be small). The underlying intuition is simple: generative models are not used to discriminate between classes (this is left to the discriminative methods), but are used to finely describe the local structure of the data as an ensemble of clusters. Even if the proposed methodology may be general (and applicable to any generative kernel), we explored this direction focusing on the HMM-based Fisher Kernel case, showing promising and comparative results obtained from some experiments. All the details may be found in [Bicego et al. 2009a] and in the D2.2 deliverable.

In the second direction of research, we defined a novel score space exploiting the free energy associated to a generative model, which is a popular score function representing a lower bound on the negative log-likelihood of the visible variables. The free energy permits to embed the uncertainty in the model parameters under the form of entropic terms. Such terms decompose in an entropy set and a cross-entropy set. The former encodes ambiguity within the model, thus considering the cases in which over-fitting or local minima occur during the learning. The latter encodes errors in the model's fit to the data, distributing such discrepancies across several terms, each one focused on a particular factor of the generative joint distribution. The entropy and cross-entropy sets are employed as features of the final score. The resulting score space shows to be highly informative for discriminative learning, allowing to achieve compelling comparative results in heterogeneous classification tasks, overwhelming the best classification accuracy on well-known databases. In particular, our approach was applied to face two

bioinformatics problems (exons/introns classification, homology detection), and to deal with a typical computer vision issue (scene/object recognition), and results are compared with the best state-of-the-art outcomes present in the literature in the respective areas. All the details may be found in the the D2.2 deliverable and the references therein, namely [Perina et al. 2009a, Perina et al. 2009b].

The third and final direction of research, mainly carried out in the last part of the second year and the following 18 months (third year and extension), was devoted to the investigation of the effect of the normalization of the score spaces. In particular we first investigated the effect of a linear normalization, focusing on the Fisher score and on the so-called trans-space (one of the score spaces introduced for HMM in the first year of the project -- see the deliverable D2.2). Actually, we have shown that a proper normalization is often essential – all details may be found in the D2.2. In the literature, this need for normalization has been also shown for other generative kernels (like the *marginalized kernel*). Nevertheless, all the employed normalizations are based on linear operations, like shifting and linear scaling; on the contrary, we investigated the usefulness of nonlinear transformations. In particular, we focus on a particular class of Score spaces, defined on generative models with latent variables (for example, the states in a Hidden Markov Model). Several nonlinear mappings are indeed possible, and we investigated different ones, based on powering operation, logarithmic and logistic functions. Some experiments on HMM-based problems assessed the validity of the proposed approach, with really promising results. All the details may be found in [Carli et al 2009] and [Carli et al 2010].

Finally, and in agreement with what was planned concerning the 6 months extension, we have developed generative embeddings specifically designed for magnetic resonance image (MRI) data, based on probabilistic models of how this type of data is generated. Specifically, given that it is known that homogenous MRI data follows a Rice distribution [Gudbjartsson and Patz, 1994], we have proposed to use a finite mixture of Riccian distributions as a model of general MRI data. Based on a Riccian mixture model estimates from training data, we define several different generative embeddings. These embeddings, when combined with the non-extensive information-theoretic kernels developed in WP2.2 (as reviewed below) yield state-of-the-art results in the problem of schizophrenia detection in MRI data. This work is described in detail in [Carli et al, 2011].

2. Task WP2.2: Compression kernels

As mentioned in the first and second year reports, this task would be better called “Information theoretic kernels”, since its goals were to devise ways to obtain kernels for non-vectorial data, based on information theoretic tools and concepts. More specifically, the idea was to assume that the objects (to be clustered or classified) were generated by some probabilistic source, and then define kernels between source models. In text categorization, this approach is an alternative to the Euclidean geometry inherent to the bag-of-words representations. In fact, approaches that map data to statistical manifolds, equipped with well-motivated non-Euclidean metrics [Lafferty and Lebanon, 2005], often outperform SVM classifiers with linear kernels [Joachims, 2002]. Some of these kernels have a natural information-

theoretic interpretation, thus bridging a gap between kernel methods and information theory [Cuturi, Fukumizu, and Vert, 2005; Hein and Bousquet, 2005].

2.1 Non-extensive Information-Theoretic Kernels

One of the main achievements of WP2.2 (mostly concluded in the first year of the project) was a new family of information-theoretic kernels, based on nonextensive information theory, which contains previous information theoretic kernels as particular elements. The famous Shannon and Rényi entropies share the so-called extensivity property: the joint entropy of two independent random variables equals the sum of their entropies. Dropping this property yields the so-called nonextensive entropies (Havrda and Charvát, 1967; Lindhard and Nielsen, 1971; Tsallis, 1988), which have raised interest among physicists for modeling phenomena with long-range interactions, and in constructing nonextensive generalizations of the Boltzmann-Gibbs statistical mechanics (Abe, 2006). The so-called Tsallis entropies (Havrda and Charvát, 1967; Tsallis, 1988) form a parametric family of nonextensive entropies that includes the Shannon entropy as a particular case. Our main achievements in this research front were the following:

(i) The concept of q-convexity, generalizing that of convexity, for which we prove a Jensen q-inequality. The related concept of Jensen q-differences, which generalize Jensen differences, is also proposed. Based on these concepts, we introduce the Jensen-Tsallis (JT) q-difference, a nonextensive generalization of the JS divergence, which is also a "mutual information" in the sense of Furuichi (2006).

(ii) Characterization of the JT q-difference, with respect to convexity and extrema, extending work by Burbea and Rao (1982) and by Lin (1991) for the JS divergence.

(iii) A broad family of (nonextensive information theoretic) positive definite kernels, interpretable as nonextensive mutual information kernels, ranging from the Boolean to the linear kernels, and including the JS kernel proposed by Hein and Bousquet (2005).

(iv) A family of (nonextensive information-theoretic) positive-definite kernels between stochastic processes, subsuming well-known string kernels (such as the well-known p-spectrum kernel) (Leslie, Eskin, and Noble, 2002).

(v) Extensions of results of Hein and Bousquet (2005) proving positive definiteness of kernels based on the unbalanced JS divergence. A connection between these new kernels and those studied by Fuglede (2005) and Hein and Bousquet (2005) is also established. In passing, we show that the parametrix approximation of the multinomial diffusion kernel (Lafferty and Lebanon, 2005) is not positive definite.

All these results on nonextensive information-theoretic kernels were described in full detail in the following publications: [Martins et al, 2008a], [Martins et al, 2008b], and [Martins et al, 2009].

2.2 Kernels from Compression Algorithms

Another important direction of work was that exploiting the use of compression algorithms as a means to approximate the information theoretic kernels, yielding the so-called compression kernels. The work on this front has mainly concentrated on practical applications of this type of kernels to a few challenging problems. In particular, we have we proposed a new compression-based ECG biometric method for personal identification and authentication, based on the Ziv-Merhav cross parsing algorithm for symbol sequences (strings), which works without any feature extraction from the waveforms. This method uses a string similarity measure obtained with a data compression algorithm and yields state-of-the-art performance both in identification and authentication tasks. We believe that this result is a clear proof of concept that compression-based (dis)similarities allow addressing difficult pattern recognition tasks, bypassing the critical feature extraction/selection step. All the details and results of this work can be found in following publications: [Pereira Coutinho et al 2010a], [Pereira Coutinho et al 2010b], [Pereira Coutinho et al 2011a], and [Pereira Coutinho et al 2011b].

2.3 Combination of Generative Embeddings with Information-Theoretic Kernels

A second direction of research exploited possible synergies in the combination of the non-extensive information theoretic kernels (namely the Jensen-Shannon and the Jensen-Tsalis kernels, and their weighted counterparts) developed in the context of WP2.2 with the generative embeddings developed in the context of WP2.1. This work involved a close collaboration between the IST and UNIVR partners, with two visits from UNIVR researchers to IST (one of which a long stay of three months of a PhD student) and resulted in several joint publications [Bicego et al 2009b], [Bicego et al 2009c], [Bicego et al, 2011], and [Figueiredo et al 2010].

To better explain the ideas that we have explored, we begin by pointing out that using a generative embedding typically involves three fundamental steps:

(i) defining and learning the generative model;

(ii) using this learned generative model to build the embedding (mapping from the original object space into a fixed-dimensional vector space, often called the *score space*);

(iii) discriminatively learning a (possibly kernel-based) classifier on the resulting score space.

The literature on generative embeddings has traditionally mainly focused on steps (i) and (ii), usually using some standard off-the-shelf tool for step (iii), such as a linear or RBF-based support vector machine or a kernel logistic regression. In our work, we have considered a different approach, by

focusing also on the discriminative learning step. In particular, we have combined the free energy score space embedding (developed in WP2.1) with the non-extensive (Tsallis) information-theoretic kernels kernels developed in the context of task WP2.2.

We have carried out experiments on a variety of classification tasks of structured objects (shapes, images), with excellent results. For example, in a scene recognition task on the benchmark Corel dataset, we have shown that this approach yields state-of-the-art performance [Bicego et al 2009c]. As reported recently in [Bicego et al, 2011], state of the art results were also obtained on three medical tasks: colon cancer detection from gene expression data, schizophrenia detection from of brain MRI images, and renal cell cancer classification from tissue microarray (TMA) data. It is important to stress that these tasks involving the MRI and TMA data are key application benchmarks of the SIMBAD project, that were the central objects of study in WP6 ("Analysis of tissue microarray (TMA) images of renal cell carcinoma") and WP7 ("Analysis of brain magnetic resonance (MR) scans for the diagnosis of mental illness").

These results seem to indicate that the non-extensive information theoretic kernels are very adequate to work on the feature spaces induced by the generative embeddings. Full details of this approach and experimental results can be found in the following publications: [Bicego et al 2009b], [Bicego et al 2009c], [Bicego et al, 2011], [Bicego et al, 2011a], and [Figueiredo et al 2010].

3. Task WP2.3: Learning and Combining Similarities

This task was devoted to the problem of learning similarities from data examples and also to combining several similarity measures. We will divide the report of this task into two parts:

(i) The first part will report our work on learning and combining similarities in the context of unsupervised learning (clustering). The following research fronts were addressed: (i) learning and combining similarities from clustering ensembles obtained under the *evidence accumulation clustering* (EAC) paradigm; (ii) learning higher order dissimilarities using the concept of *dissimilarity increments* (DI).

(ii) The second part of the report will focus on similarity learning and combination for supervised learning, with a special focus on a paradigm known as multiple kernel learning (MKL) applied to structured prediction problems.

3.1 Learning and Combining Similarities in Clustering

3.1.1 Learning and Combining Similarities from Evidence Accumulation

There is a close connection between the concepts of pairwise similarity and probability in the context of unsupervised learning. It is a common assumption that, if two objects are similar, it is very likely that they are grouped together by some clustering algorithm, the higher the similarity, the higher the

probability of co-occurrence in a cluster. Conversely, if two objects co-occur very often in the same cluster (high co-occurrence probability), then it is very likely that they are very similar. This duality and correspondence between pairwise similarity and pairwise probability within clusters forms the core idea of the *clustering ensemble* (CE) approach known as *evidence accumulation clustering* (EAC) [Fred and Jain, 2005].

Each clustering algorithm can be used to induce a pair-wise similarity. Evidence accumulation clustering combines the results of multiple clusterings into a single data partition by viewing each clustering result as an independent evidence of pairwise data organization. Using a pairwise frequency count mechanism amongst a clustering committee, the method yields, as an intermediate result, a co-association matrix that summarizes the evidence taken from the several members in the clustering ensemble. This matrix corresponds to the maximum likelihood estimate of the probability of pairs of objects being in the same group, as assessed by the clustering committee, and can be regarded as a pair-wise similarity induced by the CE. One of the main advantages of EAC is that it allows for a big diversification within the clustering committee. Indeed, no assumption is made about the algorithms used to produce the data partitions, it is robust to incomplete information, i.e., we may include partitions over sub-sampled versions of the original data set, and no restriction is made on the number of clusters of the partitions.

The EAC method can be decomposed into three major steps:

(i) construction of the clustering ensemble;
(ii) accumulation of the “clustering evidence'' provided by the ensemble;

(iii) extraction of the final consensus partition from the accumulated evidence.

Additionally, validation of the combined clustering results is a possible, sometimes desirable, final analysis step. In the following sections we outline contributions on learning similarities under the EAC framework focusing on new combination methods, constrained clustering, cluster validity, scalability issues, and applications to electrophysiological temporal data. Finally, we refer the development status of a toolbox for MATLAB, built under an object-oriented paradigm that provides an up-to-date environment for the application of the clustering ensemble approach.

3.1.2 Combining Evidence from Multiple Clusterings

The EAC approach combines evidence, from pairwise associations performed by the clustering committee, based on a voting mechanism that yields, as summarizing entity, a co-association matrix. This constitutes the intermediate step of evidence accumulation. A consensus partition is obtained by applying some clustering strategy over this matrix. Progress undertaken in the combination process was based on the dual interpretation of the co-association matrix as expressing similarities and as probabilities.

(i) Taking the pair-wise similarity, learned with the EAC method, as estimate of the probability of pairs of objects to belong to the same cluster, we proposed a probabilistic formulation for the combination process, leading to a consensus soft partition solution, where each object is probabilistically assigned to a cluster. The method reduces the clustering problem to a polynomial optimization in probability domain, which is attacked by means of the Baum-Eagon inequality. This work presents a principled probabilistic solution for consensus clustering, going one step further by extending the EAC paradigm from hard data partitioning to soft clustering solutions. More details and results can be found in [Bulò et al, 2010].

(ii) By taking co-occurrence information as the starting point, we have proposed a probabilistic generative model for consensus clustering, based on a dyadic aspect model for the evidence accumulation clustering framework, as explained in detail in [Lourenço et al, 2011]. Starting from the observation that co-occurrences are a special type of dyads, we proposed to model co-association using a generative aspect model for dyadic data. Under the proposed model, the extraction of a consensus clustering corresponds to solving a maximum likelihood estimation problem, which we address using the expectation-maximization algorithm. We referred to the resulting method as probabilistic ensemble clustering algorithm (PEnCA). Moreover, the fact that the problem is placed in a probabilistic framework allows using model selection criteria to automatically choose the number of clusters. The output of the method is a probabilistic assignment of each sample to each cluster. One of the advantages of this framework is the possibility of inclusion of a model selection criterion. We hope to further address this issue in the near future. Ongoing work on different initialization schemes and strategies to escape from local solutions is being carried out.

(iii) Different clustering techniques can be applied to the co-association matrix to obtain the combined data partition, and different clustering strategies may yield too different combination results. In an attempt to reduce the sensitivity of the final partition to this clustering method, and still obtain competitive and consistent results, we have proposed to apply embedding methods over this matrix (see [Aidos and Fred, 2011]). We performed a study of several embedding methods over the co-association matrix, interpreting it in two ways: (i) as a feature space and (ii) as a similarity space. In the first case dimensionality reduction is performed of the feature space; in the second case we obtain a new representation constrained to the similarity matrix. When applying several clustering techniques over these new representations, we evaluated the impact of these transformations in terms of performance and coherence of the obtained data partitions. Experimental results, on synthetic and real benchmark datasets, have shown that extracting the relevant features through dimensionality reduction yields more consistent results than applying the clustering algorithms directly to the co-association matrix.

The work undertaken involved a close collaboration between the IST and the UNIVE partners, which should be further strengthened in the future. In the later work, pair-wise similarities were used; extension of this method to higher order similarities is object of current joint research.

3.1.3 Constrained Clustering

Recent work on clustering has focused on the incorporation of a priori knowledge, mostly in the form of pairwise constraints, aiming to improve clustering quality of individual clustering algorithms, and find appropriate clustering solutions to specific tasks or interests.

In the context of WP2 of SIMBAD, we proposed to extend and integrate the constrained clustering idea into the CE framework. Such integration can be implemented at three main levels, and combinations thereof:

(i) on the construction of the CE, by explicitly applying constrained clustering algorithms;

(ii) during information combination phase, by forcing (hard constraints) or encouraging (soft constraints) pairwise associations;

(iii) at the step of extraction of the final (combined) data partition.

In the work developed so far, we proposed an extension to EAC (termed CEAC), and a novel algorithm (ACCCS) to solve the CE problem using pairwise constraints (must-link and cannot-link). CEAC consists of enforcing the clustering algorithm, which produces the consensus partition from the learned similarity, to support the incorporation of must-link and cannot-link constraints. The ACCCS approach comprises the maximization of both the similarity between CE partitions and a target consensus partition, and constraints satisfaction. Experimental results have shown the proposed constrained clustering combination methods performances are superior to the unconstrained EAC. Part of this work is reported in [Duarte et al, 2011].

3.1.4 Cluster Validity

Consider the following question: “for a given data set, which clustering solution should be selected?” The solution to this problem is based on clustering validation. While there is much work reported in the literature on validating data partitions produced by single clustering algorithms, little has been done in order to validate data partitions produced by clustering combination methods. Most of these works use measures of consistency between consensus solutions and the clustering ensemble, such as the Average Normalized Mutual Information proposed by [Strehl and Ghosh, 2002].

We first addressed the validation issue at the clustering ensemble level, proposing the *average cluster consistency* (ACC) index [Duarte et al, 2011]. The main idea consists of measuring how well the clusters in the clustering ensemble fit in the clusters of the consensus partition. The similarity between each partition in the CE and the combined partition is measured based on a weighting of shared samples in matching clusters. The ACC validity index accounts for the average of these similarities over the CE. Details and results on this work can be found in [Duarte et al, 2011].

In further work in this research direction, we have proposed the validation of clustering combination results at three levels:

(i) Original data representation - measure the consistency of clustering solutions with the structure of the data, perceived from the original representation (either feature-based or similarity-based);

(ii) clustering ensemble level - measure the consistency of consensus partitions with the clustering ensemble;

(iii) learned pairwise similarity - measure the coherence between clustering solutions and the co-association matrix induced by the clustering ensemble.

Taking pair-wise similarities as the underlying representation, traditional clustering validity indices (namely the Silhouette, Dunn's, and Davies and Bouldin's validity indices) were adapted to validate consensus solutions, when compared to the original data representation, and the learned similarity. These validity indices roughly account for intra-cluster compactness and inter-cluster separation. We then proposed a statistical validity index based on pair-wise similarity. According to the new index, the quality of the consensus partition is measured in terms of the likelihood of the data constrained to this partitioning. Inspired on the Parzen-window density estimation technique with variable size windows, a k-nearest neighbor density estimate from pair-wise similarities was defined. Taking as reference the learned similarity, the proposed validity index corresponds to a measure of goodness of fit of the consensus partition with the clustering ensemble and the pair-wise information extracted from it. When assessed from the original data representation, this validity index measures the goodness of fit of the combined partition with the statistical properties of the data on the baseline representation. A comparative study of the several validation approaches was undertaken on synthetic and real data. Details and results can be found in [Duarte et al, 2010].

3.1.4 Scalability

We have addressed the scalability problem of the evidence accumulation clustering method, which is intrinsically related to the storage of the co-association matrix. This topic was dealt with in collaboration with Anil K. Jain (Department of Computer Science and Engineering, Michigan State University, USA). The bottleneck of the evidence accumulation paradigm is the quadratic (on the number of samples) space complexity associated with the full representation of the co-association matrix. Taking advantage

of the sparseness of this matrix, we adopted a sparse matrix representation, reducing the space complexity of the method. In order to further reduce the space complexity, we have proposed a clustering ensemble construction rule, following a split and merge strategy, according to which the clustering algorithms are applied with K, the number of clusters, randomly chosen in the interval [Kmin,Kmax]. Criteria for the choice of these extreme values were also proposed and analyzed, showing that both space complexity and quality of combination results dependent on the partitioning granularity, dictated by the value of Kmin. Experimental results confirmed that this strategy leads to linear space complexity of evidence accumulation clustering, enabling the scalability of this framework to large data-sets. We have shown that this significant space complexity improvements do not compromise, and may even lead to increased performance of clustering combination. Details and results can be found in [Lourenço et al, 2010].

3.1.5 Learning Similarity on Temporal Data

In the context of SIMBAD, the *clustering ensemble* (CE) framework was further explored and extended to learn similarity relations of temporal data. We proposed a methodology for the analysis of data characterized by temporal evolution, such as electrophysiological signals. This methodology is based on the clustering ensemble method, and on a genetic algorithm for assessment of the existence of differentiated states in time series.

Taking as motivating application the evaluation of changes in ECG morphology in the course of the a stress-inducing computer-based activity, the evidence accumulation clustering method was applied and evaluated using different clustering algorithms for the construction of clustering ensembles as well as various algorithms for final extraction of the (combined) final partition; these various setups were additionally explored in conjunction with feature selection and feature extraction techniques. The developed work presents several innovative aspects:

(i) In previous work, stress has been found to be associated with heart rate variability. However, morphological changes have not been studied so far. In our work, we addressed this issue, by assessing the temporal evolution of ECG morphology, summarized in a similarity matrix between heart beat waves, indices of the matrix corresponding to increasing time stamps. Our results confirm this morphology change hypothesis, showing clear dissimilarity between ECG patterns at the beginning and at the end of the task; furthermore, by clustering the learned similarity matrix using the CE approach, such a hypothesis is confirmed by revealing distinct clusters.

(ii) A novel methodology for the analysis of temporal data based on the clustering ensemble approach.

(iii) Clustering of stationary temporal data with abrupt changes in the temporal organization model is a relatively simple problem. Given the continuous time evolution of stress levels,

clustering algorithms are deemed to fail to detect well separated groups of patterns. Therefore, elimination of samples that correspond to the continuous transition between distinct states (denoted as noisy patterns) is one possible approach to detect if such meaningful distinct clusters are present in the data. The genetic algorithm proposed identifies and eliminates transition time frames based on a cluster separability fitness function.

A detailed description of the previous contributions can be found in [Medina and Fred, 2010] and [Medina and Fred, 2011].

3.2 High order dissimilarities: Dissimilarity Increments

We have addressed the use of high order dissimilaritiess in pattern recognition and data mining problems. In this context, we have explored a measure of dissimilarity among triplets of nearest neighbors, called *dissimilarity increments* (DI), previously proposed in [Fred and Leitão, 2003]. In prior work, based on empirical observation, dissimilarity increments were modeled using an exponential distribution. This parametric model for cluster representation formed the basis for a new cluster isolation criterion, that was further integrated in a hierarchical clustering algorithm, having an intuitive design parameter. Within SIMBAD, we have made the following progress in this research front:

(i) We have addressed the problem of analyzing clustering solutions based on the formalism of probabilistic attributed graphs, exploring dissimilarity increments. Assuming the previously proposed statistical model for DIs, we presented a graph generative model for the clusters. This formed the basis for the design of a new cluster validity index, that consists of the description length of the data partition, represented by a probabilistic attributed graph inferred from the data, conditioned on the given partition [Fred and Jain, 2008]. Decision between clustering solutions based on the new index follows a MDL criterion. We applied the proposed criterion in two distinct scenarios: the selection of the design parameter for the hierarchical clustering algorithm mentioned above, and the choice between combination results in a clustering ensemble approach. Results on several data sets, consisting of both synthetic and real data, revealed a good performance of the index in selecting a partition or design parameter.

(ii) We have theoretically derived a statistical model of dissimilarity increments for Gaussian high-dimensional data, and have particularized the model for two

dimensional data. We empirically compared these two distributions with a prior model considered in [Fred and Leitão, 2003] (exponential distribution) using two statistical distance measures: Cramér-von-Mises and Jensen-Shannon. Empirical evidence showed that the new models provide a better approximation to the empirical distribution than the exponential distribution, while being simpler to compute [Aidos and Fred, 2011b].

(iii) We proposed the use of this distribution in clustering, having designed a novel clustering algorithm [Aidos and Fred, 2011c]: the starting point is a partition given by a Gaussian mixture decomposition and the decision of merging components is based on a likelihood ratio test between the statistical model for the combined components and the statistical model for the separate components. In [Aidos and Fred, 2011b], we proposed and evaluated another merge criterion based on the minimum description length, thus obtaining a parameter-free clustering algorithm for arbitrary shaped data, yielding state-of-the-art results in both synthetic and real-world data sets.

(iv) We have proposed to incorporate this DID in a hierarchical clustering algorithm to decide whether two clusters should be merged or not [Aidos and Fred, 2011c]: the novel hierarchical algorithm is parameter-free and can identify classes as the union of clusters following the dissimilarity increments distribution. Experimental results have shown that the proposed algorithm has excellent performance over well separated clusters, also providing a good hierarchical structure insight into touching clusters.

(v) We have presented a novel maximum a posteriori (MAP) classifier which uses the dissimilarity increments distribution. This classifier, which we named MAP-DID, can be interpreted as a Gaussian mixture model with a ``harmonizing'' operator which forces a class to have a common increment structure, even though each gaussian component within a class can have distinct means and covariances. We have applied the classifier to the dissimilarity data sets assembled in WP3, both in their original dissimilarity representations, and over the several embeddings therein explored. We have shown that MAP-DID outperforms multiple other classifiers on the various datasets (both synthetic and real) and embedding feature spaces. This work is reported in [Aidos et al, 2011d].

Although theoretically derived for Gaussian data, we have shown empirically that application of the dissimilarity increment distribution to arbitrary data sets, without the constraint of Gaussianity, leads to good performances, both under the supervised and unsupervised

approaches. Ongoing work includes a more general theoretical derivation of the distribution of dissimilarity increments, with no assumption about the generating model, focusing on the distribution for the nearest neighbors. Our future plans involve further exploration of the dissimilarity increments for classification purposes. This work is being conducted in collaboration with TU-Delft.

3.3 Supervised Learning: Multiple Kernel Learning

Despite all the advances in kernel-based machine learning, obtaining good predictors still requires a large effort in feature/kernel design and tuning (often done via cross-validation). Because discriminative training of structured predictors can be quite slow, especially in large-scale settings, it is appealing to learn the kernel function simultaneously. In multiple kernel learning (MKL, see [Lanckriet et al 2004], [Bach et al, 2004]), the kernel is learned as a linear combination of prespecified base kernels. This framework has been made scalable with the advent of wrapper-based methods, in which a standard learning problem (for example, an SVM) is repeatedly solved in an inner loop up to a prescribed accuracy [Sonnenburg et al 2006], [Zien and Ong, 2007], [Rakotomamonjy et al, 2008]. Unfortunately, extending such methods to large-scale (namely, structured prediction) still raises practical hurdles: when the output space is large, so are the kernel matrices, and the number of support vectors; when it is prohibitive to tackle the inner learning problem in its batch form, one often needs to resort to online algorithms [Ratliff et al, 2006], [Collins et al, 2008], [Shalev-Shwartz et al, 2007]; the latter are fast learners but slow optimizers [Bottou and Bousquet, 2007], hence using them in the inner loop with early stopping may misguide the overall MKL optimization.

In our work in this context, we have proposed to overcome the above difficulties by proposing a stand-alone online MKL algorithm, which exploits the large-scale tradeoffs directly. The algorithm, which when applied to structured prediction problems is termed SPOM (*Structured Prediction by Online MKL*), iterates between subgradient and proximal steps, and has important advantages:

(i) it is simple, flexible, and compatible with both sparse and non-sparse variants of MKL;

(ii) it is adequate for large-scale structured prediction problems;

(iii) it offers regret, convergence, and generalization guarantees. In fact, our approach can be seen as a kernelization (similarity-based-version) of the recent forward-

backward splitting scheme *Fobos* [Duchi and Singer, 2009], whose regret bound we improve.

This work, which has a with a special emphasis on its application to large-scale structured prediction problems, was reported in detail in an AISTAST'2011 (April, 2011) paper [Martins et al, 2011] and, previously, in a paper presented at the *NIPS Workshop in New Directions in Multiple Kernel Learning* [Martins et al, 2010].

Publications Resulting for WP2

[Aidos and Fred, 2011] H. Aidos and A. Fred, "A study of embedding methods under the evidence accumulation framework", *Proceedings of the 1st International Workshop on Similarity-Based Pattern analysis and Recognition – SIMBAD'2011*, Springer, 2011 (to appear).

[Aidos and Fred, 2011b] H. Aidos and A. Fred, "Statistical modeling of dissimilarity increments for d-dimensional data: application in partitional clustering", submitted to *Pattern Recognition*, 2011.

[Aidos and Fred, 2011c] H. Aidos and A. Fred, "On the distribution of dissimilarity increments", in J. Vitrià, J. Sanches, and M., Hernández, editors, *Pattern Recognition and Image Analysis*, LNCS vol. 6669, pp. 192–199, Springer, 2011.

[Aidos et al, 2011d] H. Aidos, A. Fred, and B. Duin, "Classification using high order dissimilarities in non-Euclidean spaces". Submitted to *1st International Conference on Pattern Recognition Applications and Methods – ICPRAM'2012,* 2011.

[Bicego et al, 2009a]: M. Bicego, M. Cristani, V. Murino, E. Pekalska, R. Duin, "Clustering-based construction of Hidden Markov Models for generative kernels", *Proc. of Int. Conf. on Energy Minimization Methods in Computer Vision and Pattern Recognition (EMMCVPR2009),* pp. 466-479, Bonn, Germany, 2009.

[Bicego et al, 2009b] M. Bicego, A. Martins, V. Murino, P. Aguiar, and M. Figueiredo, "2D shape recognition using information theoretic kernels", International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010.

[Bicego et al, 2009c] M. Bicego, A. Perina, V. Murino, A. Martins, P. Aguiar, and M. Figueiredo, "Combining free energy score spaces with information theoretic kernels: application to scene classification", International Conference on Image Processing (ICIP), Hong Kong, China, 2010.

[Bicego et al, 2011] M.Bicego, A. Ulaş, U. Castellani, A. Perina, V. Murino, A. Martins, P. Aguiar, and M. Figueiredo, "Combining information theoretic kernels with generative embeddings for classification", submitted to *Pattern Recognition*, 2011.

[Bicego et al, 2011a] M. Bicego, A. Ulaş, P. Schüffler, U. Castellani, P. Mirtuono, V. Murino, A. Martins, P. Aguiar, M. Figueiredo, "Renal cancer cell classification using generative embeddings and information theoretic kernels", *IAPR International Conference on Pattern Recognition in Bioinformatics - PRIB 2011*, Delft, The Netherlands, 2011.

[Bulò et al, 2010] S. Bulò, A. Lourenço, A. Fred, and M. Pelillo, "Pairwise probabilistic clustering using evidence accumulation", in E. Hancock, R. Wilson, T. Windeatt, I. Ulusoy, and F. Escolano, editors,

Structural, Syntactic, and Statistical Pattern Recognition, LNCS volume 6218, pp. 395-404, Springer, 2010.

[Carli et al, 2009] A. Carli, M. Bicego, S. Baldo, and V. Murino, "Non-linear Generative embeddings for kernels on latent variable models", *Proceedings of the ICCV'2009 Workshop on Subspace Methods*, Kyoto, Japan, 2009.

[Carli et al, 2010] A. Carli, M. Bicego, S. Baldo, and V. Murino, "Nonlinear mappings for generative kernels on latent variable models", Proceedings of the International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010.

[Carli et al, 2010] A. Carli, M. Figueiredo, M. Bicego, and V. Murino, "Generative Embeddings Based on Rician Mixtures: Application to Kernel-Based Discriminative Classification of Magnetic Resonance Images", Submitted to *1st International Conference on Pattern Recognition Applications and Methods – ICPRAM'2012,* 2011.

[Duarte et al, 2011] F. Duarte, J. Duarte, A. Fred, and M. Rodrigues, "Average cluster consistency for cluster ensemble selection", in A. Fred, J. Dietz, K. Liu, and J. Filipe, editors, *Knowledge Discovery, Knowlege Engineering and Knowledge Management*, Communications in Computer and Information Science, volume 128, Springer, 2011.

[Duarte et al, 2010] J. Duarte, A. Fred, A. Lourenço, and F. Duarte, "On consensus clustering validation." In E. Hancock, R. Wilson, T. Windeatt, I. Ulusoy, F. Escolano, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, LNCS vol. 6218, pp. 385-294, Springer, 2010.

[Figueiredo et al, 2010] M. Figueiredo, P. Aguiar, A. Martins, V. Murino, and M. Bicego, "Information theoretical kernels for generative embeddings based on hidden Markov models", In E. Hancock, R. Wilson, T. Windeatt, I. Ulusoy, F. Escolano, editors, *Structural, Syntactic, and Statistical Pattern Recognition*, LNCS vol. 6218, pp. 395-404, Springer, 2010.

[Fred and Jain, 2008] A. Fred and A. Jain, "Cluster validation using a probabilistic attributed graph", *Proceedings of the 19th International Conference on Pattern Recognition – ICPR'2008*, pp. 2360-2363, Tampa, Florida, USA, 2008.

[Lourenço et al, 2010] A. Lourenço, A. Fred, and A. Jain, "On the scalability of evidence accumulation clustering." *Proceedings of the 20th International Conference on Pattern Recognition – ICPR'2010*, Istanbul, Turkey, 2010.

[Lourenço et al, 2011] A. Lourenço, A. Fred, and M. Figueiredo, "A generative dyadic aspect model for evidence accumulation clustering" *Proceedings of the 1st International Workshop on Similarity-Based Pattern analysis and Recognition – SIMBAD'2011*, Springer, 2011 (to appear).

[Medina and Fred, 2010] L. Medina, A. Fred. "Clustering temporal data: application to electrophysiological signals", *Communications in Computer and Information Science*, Springer, 2010.

[Medina and Fred, 2011] L. Medina and A. Fred, "Clustering data with temporal evolution: Application to electrophysiological signals". In J. Filipe, A. Fred, B. Sharp, editors, *Agents and Artificial Intelligence*, Communications in Computer and Information Science, volume 129, Springer, 2011.

[Martins et al, 2008a] A. Martins, P. Aguiar, and M. Figueiredo. “Tsallis kernels on measures” , Proceedings of the *IEEE Information Theory Workshop*, Porto, Portugal, 2008.

[Martins et al, 2008b] A. Martins, M. Figueiredo, P. Aguiar, N. Smith, and E. Xing. “Nonextensive entropic kernels”, *Proceedings of the International Conference on Machine Learning – ICML’08*, Helsinki, inland, 2008.

[Martins et al, 2009] A. Martins, N. Smith, E. Xing, P. Aguiar, and M. Figueiredo, “Nonextensive information-theoretic kernels on measures”, *Journal of Machine Learning Research*, vol. 10, pp. 935-975, 2009.

[Martins et al, 2010] A. Martins, N. Smith, E. Xing, P. Aguiar, M. Figueiredo, “Online MKL for Structured Prediction", *NIPS2010 Workshop on New Directions in Multiple Kernel Learning*, Whistler, Canada, 2010.

[Martins et al, 2011] A. Martins, N. Smith, E. Xing, P. Aguiar, M. Figueiredo, “Online learning of tructured predictors with multiple kernels", *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics - AISTATS'2011*, Fort Lauderdale, FL, USA, 2011.

[Pereira Coutinho et al, 2010a] D. Pereira Coutinho, A. Fred, M. Figueiredo, “'Personal identification and authentication based on one-lead ECG by using Ziv-Merhav cross parsing”, 10th International Workshop on Pattern Recognition in Information Systems (PRIS), Funchal, Portugal, 2010.

[Pereira Coutinho et al, 2010b] D. Pereira Coutinho, A. Fred, M. Figueiredo, “One-lead ECG-based personal identification using Ziv-Merhav cross parsing”, Proceedings of the International Conference on Pattern Recognition (ICPR), Istanbul, Turkey, 2010.

[Pereira Coutinho et al, 2011a] D. Pereira Coutinho, A. Fred, and M. Figueiredo, “ECG-based continuous authentication system using adaptive string matching”, *Proceedings of the International Conference on Bio-inspired Systems and Signal Processing -BIOSIGNALS*, Rome, Italy, 2011.

[Pereira Coutinho et al, 2011b] D. Pereira Coutinho, H. Silva, H. Gamboa, A. Fred, and M. Figueiredo, “A case study on fiducial and non-fiducial approaches to ECG-based biometric systems.” Submitted to *Pattern Analysis and Applications*, 2011.

[Perina et al, 2009a] A. Perina, M. Cristani, U. Castellani, V. Murino, and N.Jojic. “Free energy score space”, *Advances in Neural Information Processing Systems (NIPS)*, Vancouver, Canada, 2009.

[Perina et al, 2009b] A. Perina, M. Cristani, U. Castellani, V. Murino, and N.Jojic. A hybrid generative/discriminative classification framework based on free-energy terms”, *Proceedings of the International Conference on Computer Vision (ICCV),* Kyoto, Japan, 2009.

[Lourenço et al, 2010] A. Lourenço, A. Fred, A. K. Jain, “On the scalability of evidence accumulation clustering”, *Proceedings of the International Conference on Pattern Recognition (ICPR),* Istanbul, Turkey, 2010.

Other references cited in this report:

[Bach et al, 2004] F. Bach, G. Lanckriet, and M. Jordan, "Multiple kernel learning, conic duality, and the SMO algorithm", *Proceedings of the International Conference on Machine Learning – ICML*, 2004.

[Bottou and Bousquet, 2007] L. Bottou and O. Bousquet, "The tradeoffs of large scale learning", *Neural Information Processing Systems - NIPS*, 2007.

[Burbea and Rao, 1982] J. Burbea and C. Rao, "On the convexity of some divergence measures based on entropy functions", *IEEE Transactions on Information Theory*, vol. 28, pp. 489–495, 1982.

[Collins et al, 2008] M. Collins, A. Globerson, T. Koo, X. Carreras, and P. Bartlett, "Exponentiated gradient algorithms for conditional random fields and max-margin Markov networks", *Journal of Machine Learning Research,* vol. 9*,* pp. 1775-1822, 2008.

[Cuturi, Fukumizu, and Vert, 2005] M. Cuturi, K. Fukumizu, and J.-P. Vert. "Semigroup kernels on measures", *Journal of Machine Learning Research*, vol. 6, pp. 1169–1198, 2005.

[Furuichi, 2006] S. Furuichi, "Information theoretical properties of Tsallis entropies", *Journal of Mathematical Physics*, vol. 47, no. 2, 2006.

[Fred and Jain, 2005] A. Fred and A. Jain, "Combining multiple clusterings using evidence accumulation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, pp. 835-850, 2005.

[Fred and Leitão, 2003] A. Fred, and J. Leitão, "A new cluster isolation criterion based on dissimilarity increments", *IEEE Trans. on Pattern Analysis and Machine Intelligence,* vol. 25, pp. 944-958, 2003.

[Gudbjartsson and Patz, 1994] H. Gudbjartsson and S. Patz, "The Rician distribution of noisy MRI data", *Magnetic Resonance in Medicine*, vol. 34, pp. 910-914, 1994.

[Havrda and Charvát, 1967] M. Havrda and F. Charvat. "Quantification method of classification processes: concept of structural α-entropy", *Kybernetika*, vol. 3, pp. 30–35, 1967.

[Hein and Bousquet, 2005] M. Hein and O. Bousquet, "Hilbertian metrics and positive definite kernels on probability measures", in Z. Ghahramani and R. Cowell, editors, *Proceedings of the 10th International Workshop on Artificial Intelligence and Statistics (AISTATS)*, 2005.

[Joachims, 2002] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory and Algorithms*. Kluwer Academic Publishers, 2002.

[Lafferty and Lebanon, 2005] J. Lafferty and G. Lebanon, "Diffusion kernels on statistical manifolds", *Journal of Machine Learning Research*, vol. 6, pp. 129–163, 2005.

[Lanckriet et al, 2004] G. Lanckriet, N. Cristianini, P. Bartlett, L. Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming", *Journal of Machine Learning Research*, vol. 5, pp. 27-72, 2004.

[Leslie, Eskin, and Noble, 2002] C. Leslie, E. Eskin, and W. Noble, "The spectrum kernel: A string kernel for SVM protein classification", *Proceedings of the Pacific Symposium on Biocomputing*, pp. 564–575, 2002.

[Lin, 1991] J. Lin, "Divergence measures based on the Shannon entropy", *IEEE Transactions on Information Theory*, vol. 37, pp. 145–151, 1991.

[Lindhard and Nielsen, 1971] J. Lindhard and V. Nielsen, *Studies in Statistical Dynamics*, Munksgaard, Copenhagen, 1971.

[Rakotomamonjy et al, 2008] A. Rakotomamonjy, F. Bach, S. Canu, and Y. Grandvalet, "SimpleMKL", *Journal of Machine Learning Researc,* vol. 9, pp. 2491-2521, 2008.

[Ratliff et al, 2006] N. Ratliff, J. Bagnell, and M. Zinkevich, "Subgradient methods for maximum margin structured learning", *ICML Workshop on Learning in Structured Outputs Spaces*, 2006.

[Shalev-Shwartz et al, 2007] S. Shalev-Shwartz, Y. Singer, and N. Srebro, "Pegasos: Primal estimated sub-gradient solver for SVM", *Proceedings of the International Conference on Machine Learning – ICML,* 2007.

[Sonnenburg et al, 2006] S. Sonnenburg, G. Rätsch, C. Schäfer, and B. Schölkopf, "Large scale multiple kernel learning", *Journal of Machine Learning Research,* vol. 7, pp. 1531-1565, 2006.

[Strehl and Ghosh, 2002] A. Strehl and J. Ghosh, "Cluster Ensembles { A Knowledge Reuse Framework for Combining Multiple Partitions", *Journal of Machine Learning Research*, vol. 3, pp. 583-617, 2002.

[Tsallis, 1988] C. Tsallis, "Possible generalization of Boltzmann-Gibbs statistics", *Journal of Statistical Physics*, vol. 52, pp. 479–487, 1988.

[Zien and Ong, 2007] A. Zien and C. Ong, "Multiclass multiple kernel learning", *Proceedings of the International Conference on Machine Learning – ICML*, 2007.