Learning with general proximity measures

Elżbieta Pękalska[1,2] and Robert P.W. Duin[1]

[1]Faculty of Electrical Engineering, Mathematics and Computer Sciences,
Delft University of Technology, The Netherlands
[2]School of Computer Science, University of Manchester, United Kingdom
e.m.pekalska@tudelft.nl

Abstract. Proximity is the basic quality which identifies and characterizes groups of objects in various domains and contexts. When objects are compared to a set of chosen prototype examples, proximity can be used as a natural ingredient to build a numerical representation. Pattern classes may be learned from such proximity representations by the traditional nearest neighbor rule, as well as by other alternative strategies. These encode the proximity information in suitable representation vector spaces in which statistical classifiers can be trained. Such recognition techniques can be successful, provided that the measure is informative, independently whether it is metric or Euclidean, or not.

1 Introduction

In our attempts to model the physical world and its phenomena we usually represent objects as points with masses in a Euclidean space. While it is convenient to use, we should be aware of the imposed simplification in our thinking. If we treat objects as bodies in the space, the actual distance, i.e. the shortest distance between them is non-metric. For instance, a zero distance occurs between two touching objects, even if they are very different; two mugs on a table touch the table and have, thereby, zero distances to it, while they can be far away from each other. In daily life we often deal asymmetric proximity measures, which result from different judgements used to view one object in the context of another. A practical example are asymmetric distances between places in a city, where asymmetry is caused by one-way road connections. Thanks to the well-developed theory and methods of vector spaces, the Euclidean distance has widely been accepted as (one of) the most suitable measures to rely on in building our models and explaining the world. The reason is that a Euclidean space is simultaneously an inner product, normed and metric space. While this offers an arsenal of powerful approaches, we should not forget that a vectorial representation is a reduction of complex phenomena.

Statistical learning is one of the primary approaches to pattern recognition. It assumes that knowledge can be represented in a set of discriminative features. Hence, objects are represented as vectors in a feature space, which is usually Euclidean. The proximity between objects is modeled as the Euclidean distance between their vector representations. Learning relies on inferring a functional relation between some input and output data in this space, given a set of examples. This is done to assure its high predictive ability.

On the other hand, knowledge is usually qualitative in nature. Moreover, many problems deal with structural data descriptions which are non-vectorial by origin. The most basic quality to be used in learning is therefore the observation of how much the objects have in common or how much they differ. As a result, a suitable proximity measure can be designed in a given context in order to identify patterns and model clusters. Such a measure becomes powerful, if it is derived by focussing on differences occurring in the structure of objects. Consequently, objects can be represented by a vector of proximity values to some chosen prototypes. Since expert knowledge can be used in the definition of the measure, such proximity representations enable the natural integration of qualitative knowledge with numerical methodology. They combine the strengths of structural and statistical approaches: a structural measure builds a numerical representation, which is used in statistical learning. Note that kernel methods are mathematically elegant approaches [1] which lie within the proximity-based paradigm. They rely on a relatively narrow class of (conditionally) positive semidefinite kernels and are specific types of similarity representations.

Many natural proximity measures used to compare contours, sequences, spectra or images are non-metric or do not possess the Euclidean behavior. Non-metric examples are pairwise structural alignments of proteins that focus on local similarity [2,3], variants of the Hausdorff distance [4], the normalized edit distance [5], the Mahalanobis distance between populations [6] or the Kullback-Leibler divergence [7]. The violation of metric axioms is often not an artifact of poor choice of primitives, features or algorithms, but inherent to the problem of a proper comparison of objects that incorporates the necessary invariance and is robust against noise or occlusion [8].

Although proximity measures are widely used for matching and object comparison [5,4,8,9], classification often relies on assigning a new object to the class of its nearest neighbor. Alternative generalization frameworks, however, exist for general proximity measures. They represent proximity information in suitable representation vector spaces [10,11,3,6] or deal with indefinite kernels [12,13,6].

In this paper we present a brief description of proximity-based statistical learning approaches. Since many measures defined in practice are non-Euclidean or non-metric, we emphasize the necessity of developing novel learning strategies that make use of informative aspects of the measure and not necessarily of its metric properties. Our claim is that metric or Euclidean requirements are not essential if the measure is discriminative.

2 Learning from proximity data

Assume a representation set $R = \{p_1, p_2, \ldots, p_n\}$ of prototype examples and a proximity measure d, which should incorporate the necessary invariance. Without loss of generality, let d denote dissimilarity. An object x is then represented as a vector of dissimilarities computed between x and the prototypes from R, i.e. $d(x, R) = [d(x, p_1), d(x, p_2), \ldots, d(x, p_n)]^{\mathsf{T}}$. Given a set $T = \{t_1, t_2, \ldots, t_N\}$ of N objects, our proximity representation becomes an $N \times n$ dissimilarity matrix $D(T, R)$, where $D(t_i, R)$ is now a row vector. If $R \subset T$, then R is usually selected

out of T in a way to guarantee a good tradeoff between the recognition accuracy and the computational complexity.

The k-NN rule can directly be applied to pairwise proximity data. Although it has good asymptotic properties for metric distances, its performance deteriorates for small training (representation) sets. Alternative learning strategies represent proximity information in suitable representation vector spaces, in which traditional statistical algorithms can be defined. So, they can become more beneficial. Two simple approaches are a linear isometric embedding into a pseudo-Euclidean space and the use of dissimilarity spaces.

Pseudo-Euclidean linear embedding. Given a symmetric dissimilarity matrix $D(R,R)$, a vectorial representation X can be found such that the distances are preserved. This is done in a pseudo-Euclidean space $\mathcal{E}=\mathbb{R}^{(p,q)}$, which is a $(p+q)$-dimensional non-degenerate indefinite inner product space such that the inner product $\langle\cdot,\cdot\rangle_{\mathcal{E}}$ is positive definite on $\mathbb{R}^p$ and negative definite on $\mathbb{R}^q$ [14]. Then, $\langle\mathbf{x},\mathbf{y}\rangle_{\mathcal{E}}=\mathbf{x}^T\mathcal{J}_{pq}\mathbf{y}$, where $\mathcal{J}_{pq}=\mathrm{diag}\,(I_{p\times p};-I_{q\times q})$ and I is the identity matrix. Consequently, $d^2_{\mathcal{E}}(\mathbf{x},\mathbf{y})=\langle\mathbf{x}-\mathbf{y},\mathbf{x}-\mathbf{y}\rangle_{\mathcal{E}}=d^2_{\mathbb{R}^p}(\mathbf{x},\mathbf{y})-d^2_{\mathbb{R}^q}(\mathbf{x},\mathbf{y})$, hence $d^2_{\mathcal{E}}$ is a difference of square Euclidean distances found in $\mathbb{R}^p$ and $\mathbb{R}^q$. Since $\mathcal{E}$ is a linear space, many properties related to inner products can be extended from the Euclidean case [14,6].

The (indefinite) Gram matrix G of X can be expressed by the square distances $D^{\star 2}=(d^2_{ij})$ as $G=-\frac{1}{2}JD^{\star 2}J$, where $J=I-\frac{1}{r}\mathbf{1}\mathbf{1}^T$ [14,10,6]. Hence, X can be determined by the eigendecomposion of $G=Q\Lambda Q^T=Q|\Lambda|^{1/2}\mathrm{diag}(\mathcal{J}_{p'q'};0)\,|\Lambda|^{1/2}Q^T$, where $|\Lambda|$ is a diagonal matrix of first decreasing p' positive eigenvalues, then decreasing magnitudes of q' negative eigenvalues, followed by zeros. Q is a matrix of the corresponding eigenvectors. X is uncorrelated and represented in $\mathbb{R}^k$, $k=p'+q'$, as $X=Q_k|\Lambda_k|^{1/2}$ [14,10]. Since only some eigenvalues are significant (in magnitude), the remaining ones can be disregarded as non-informative. The reduced representation $X_r=Q_m\,|\Lambda_m|^{1/2}$, $m=p+q<k$, is determined by the largest p positive and the smallest q negative eigenvalues. New objects $D(T_{test},R)$ are orthogonally projected onto $\mathbb{R}^m$; see [14,10,6] for details. Classifiers based on inner products can appropriately be defined in $\mathcal{E}$. A linear classifier $f(\mathbf{x})=\mathbf{v}^T\mathcal{J}_{pq}\mathbf{x}+v_0$ is e.g. constructed by addressing it as $f(\mathbf{x})=\mathbf{w}^T\mathbf{x}+v_0$, where $\mathbf{w}=\mathcal{J}_{pq}\mathbf{v}$ in the associated Euclidean space $\mathbb{R}^{(p+q)}$ [14,10,6].

Proximity spaces. Here, the dissimilarity matrix $D(X,R)$ is interpreted as a data-dependent mapping $D(\cdot,R)$: $X\to\mathbb{R}^n$ from some initial representation X to a vector space defined by the set R. This is the *dissimilarity space*, in which each dimension $D(\cdot,p_i)$ corresponds to a dissimilarity to a prototype $p_i\in R$. The property that dissimilarities should be small for similar objects (belonging to the same class) and large for distinct objects, gives them a discriminative power. Hence, $D(\cdot,p_i)$ can be interpreted as 'features' and traditional statistical classifiers can be defined [3,6]. Although the classifiers are trained on $D(\cdot,R)$, the weights are still optimized on the complete set T. Thereby, they can outperform the k-NN rule as they become more global in their decisions.

Normal density based classifiers perform well in dissimilarity spaces [10,3,6]. For a two-class problem, a quadratic normal density based classifier (NQC) is given by $f(D(x,R)) = \sum_{i=1}^{2} \frac{(-1)^i}{2}(D(x,R)-\mathbf{m}_i)^\mathsf{T} S_i^{-1}(D(x,R)-\mathbf{m}_i) + \log\frac{p_1}{p_2} + \frac{1}{2}\log\frac{|S_1|}{|S_2|}$, where $\mathbf{m}_{1/2}$ are the mean vectors and $S_{1/2}$ are the estimated class covariance matrices computed in the dissimilarity space $D(X,R)$. $p_{1/2}$ are the class prior probabilities. If $S_{1/2}$ are replaced by the average covariance matrix, then a linear classifier is obtained. If the covariance matrices become singular, they need to be regularized. Here, we choose the following regularization $S_i^\kappa = (1-\kappa)S_i + \kappa p_i \operatorname{diag}(S_i)$, $\kappa \in [0,1]$, which leads to the regularized NQC (RNQC).

Another useful strategy is the class of sparse linear programming machines (LPMs), which construct hyperplanes in the corresponding dissimilarity spaces. They are able to automatically determine a prototype set R (or if trained on $D(T,R)$, they may reduce R further on) which defines the final classifier. Two variants are considered: the μ-LPM and the auc-LPM. The μ-LPM is a form of the ℓ_1-SVM with $\mu \in [0,1)$ being related to the expected classification error [15,3]. The auc-LPM is defined to maximize the area under the ROC curve, as recently proposed in [16]. The LPMs are trained on a complete representation $D(T,T)$ and determine both R and the weights of the classifiers. Additionally, the NLSQC (nonnegative least square classifier) is used [17]. It is a linear function which optimizes a square error and is, thereby, competitive to a quadratic classifier. It has no additional parameters. The NLSQC may compete with other LPMs applied to dissimilarity data, but its solution is not very sparse in terms of R. So, the sets R found by the LPMs may be used to train the NLSQC on $D(T,R)$ to enhance its sparsity.

3 Clustering example

As an illustration of clustering we consider a subset of $n = 70$ fish contours[1]. They are first aligned with the main axes and normalized to the same bounding box. Then, they are re-sampled to 40 equidistant points. Assuming we know the point-to-point correspondences between pairs of contours, Procrustes analysis is applied to derive the distance [18]. In this process, one of the contours is kept fixed, while the other is transformed by some optimal scaling, shift and rotation. The measure of fit is based on the sum of Euclidean distances between the corresponding points (the Frobenius norm between 40×2 matrices representing the contours). Since the fish contours can be arbitrary positioned, we consider all cyclic correspondences and determine the one with the smallest mismatch. Finally, such a distance is scaled to [0,1] when normalized by the Frobenius norm of the first contour. This becomes our symmetric dissimilarity representation D.

To order to emphasize possible clusters, a new matrix $D*$ is computed such that $d^*(i,j) = \frac{2n\, d(i,j)}{\sum_k d(i,k) + \sum_l d(l,j)}$. Finally, as we wish to look for grouping tendencies, we define a similarity representation $S^* = d_m^* \mathbf{1}\mathbf{1}^T - D^*$, where d_m^* is the maximal element in D^*. Three clustering techniques are considered. These are

[1] http://www.ee.surrey.ac.uk/Personal/F.Mokhtarian/

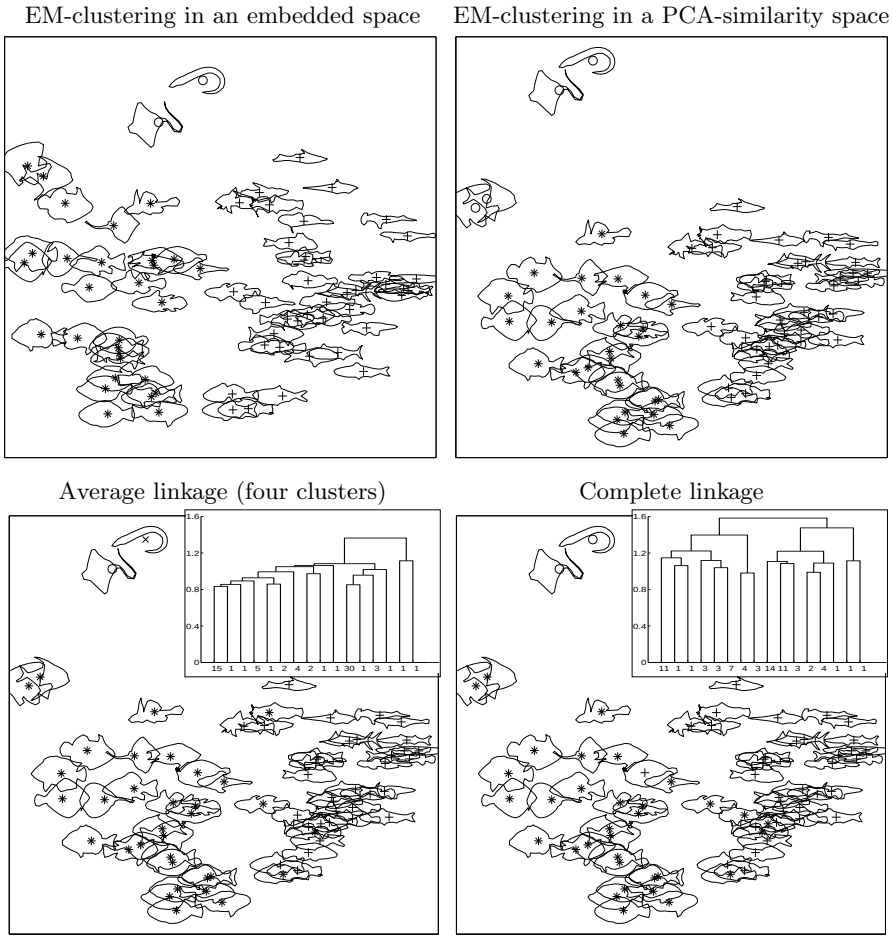

Fig. 1. Clustering results for Fish contours. Clusters are denoted by different marks: '+','*' and 'o'. In the top row, the plots show the 2D embedded space (left) and the PCA-similarity (right) spaces with the projected fish contours. (The EM-clustering is performed in higher-dimensional spaces). The axes are flipped such that the results can be compared. In the bottom row, the 2D PCA-similarity space is used to represent the clustering results for the average linkage (left) and complete linkage (right) . Partial dendrograms are presented in the right corners of these plots, indicating the number of objects in each sub-cluster on the horizontal axis.

the hierarchical clustering on D^* and the EM-clustering applied in two representation spaces: a pseudo-Euclidean embedded space and a PCA-transformed similarity space (with 90% preserved variance), both determined from S^*. The embedded space is found by assuming that the Gram matrix discussed in Sec. 2 is defined as $G = \frac{1}{2}JS^*J$. Its dimensionality is automatically detected by the number of significant eigenvalues. This procedure resembles spectral clustering [19], except that the scaling is different and the dimensionality of our embedded space may be higher.

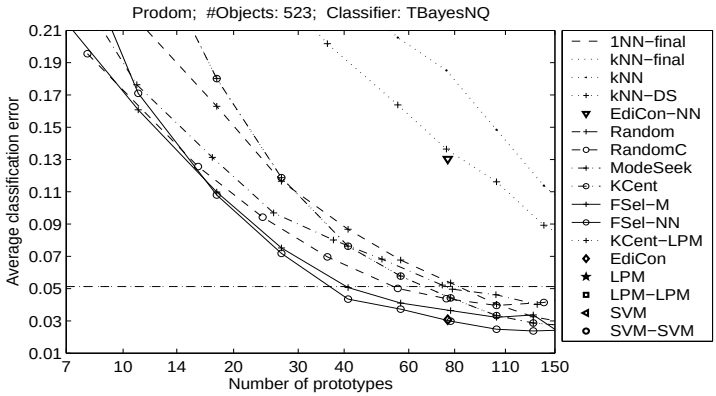

Fig. 2. Four-class ProDom data. Averaged (over 30 runs) classification error of various dissimilarity-based classifiers as a function of the number of selected prototypes. Standard deviations of the average errors are not shown to maintain the clarity. They are less than 0.004. The LPM and the SVM achieve small errors, but they require 400 prototypes, and are not shown on the plot. Note also that the x-axis is logarithmic.

There are possibly two clusters in the data (as observed from the number of 'standing-out' eigenvalues of G), but they are not very apparent. As some 'outliers' exist, we set the number of clusters to three. As there are weak grouping tendencies based on the object-to-objects distances, single linkage gives bad results. Complete linkage is able to detect three clusters, as well as average linkage, however, the clusters are weak. The EM-clustering in an embedded space and in the PCA-similarity space give better results, as shown in Fig.1. As these methods depend on initialization, they are run 30 times and the best results are chosen according to the maximum of the goodness-of-clustering criterion, $J_{GOC} = \frac{\sum_i ni/(n-n_i)\sum_j n_j A_{ij}}{2\sum_i n_i A_{ii}}$. n_i is the cluster cardinality, A_{ij} is the average dissimilarity between the i-th and j-th clusters.

4 Classification examples

We will first present some indices characterizing proximity data. Assume K classes, $\omega_1, \ldots, \omega_K$ such that $|\omega_i| = n_i$ and $N = \sum_i n_i$. The class separability is defined as $J_{\text{sep}} = \frac{1}{J_{GOC}}$. The smaller the value, the better the separability. Concerning the deviation from the Euclidean behavior, a symmetric $D(T,T)$ has a Euclidean behavior iff the Gram matrix $G = -\frac{1}{2} J D^{\star 2} J$ is positive semidefinite [10,6]. It means that all eigenvalues λ_i of G are nonnegative. Hence, the magnitudes of negative eigenvalues manifest the amount of deviation from the Euclidean behavior. This is presented in the indices $J_{\text{eigM}} = |\lambda_{min}|/\lambda_{max}$, i.e. the ratio of the absolute value of the smallest negative eigenvalue to the largest positive one, and $J_{\text{eigS}} = \sum_{\lambda_i<0} |\lambda_i| / \sum_{j=1}^{N} |\lambda_j|$, i.e. the overall contribution of negative eigenvalues. Concerning non-metric aspects, any symmetric D can be made metric by adding a suitable constant γ to all off-diagonal elements of D. In a first attempt, it can be overestimated by $\gamma_0 = \max_{p,q,t} |d_{pq} + d_{pt} - d_{qt}|$ [6]. Given that, a better estimation of $\gamma \in (0, \gamma_0)$ is found by an iterative bisection

method. Our index is, therefore, $J_\gamma = \gamma \geq 0$. Another way to characterize the non-metric behavior is by the number of disobeyed triangle inequalities J_{ineq}.

ProDom data. A *ProDom* subset of 2604 protein domain sequences from the ProDom set [20] is considered, together with the pairwise structural alignments s_{ij}, as defined in [2]. It is a four-class problem: 878/404/271/1051 examples. The dissimilarities are derived as $d_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}}/\sqrt{2604}$. The average distance in the training data equals to 11.8 and the maximum distance is 18.5, while $J_{\text{eigS}} = 0.0048$ and $J_{\text{eigM}} = 0.009$, $J_{\text{ineq}} \leq 6$ and $J_\gamma \approx 0.173$. Consequently, the measure is nearly metric and nearly Euclidean.

Due to lack of space, we focus only on the dissimilarity space approach. In our experiments, the data set is randomly divided into a training set T, $|T| = 523$ and a test set S. Then, a representation set R is chosen from T according to some criteria. We choose to train the RQNC in a dissimilarity space $D(T, R)$ and test it on $D(S, R)$. The regularization parameter κ is determined in a 5-fold cross-validation. This is repeated 30 times and the results are averaged.

Except for random selection (Random, or random per class, RandomC), many other selection procedures can be considered, e.g. by suitable adaptations of clustering approaches, such as the k-centers (KCent) or mode-seek (ModeSeek) clustering. Briefly, they look for objects that are either local centers of local modes in the dissimilarity data [3]. Concerning dissimilarity spaces, supervised approaches include editing and condensing (EdCon) and the sparse μ-LPM. Since there are four unbalanced classes, the sparsity of the μ-LPM may not be large, so we also apply the k-centers (to pre-select the representation set) followed by the LPM (KCent-LPM). We also run the indefinite support vector machine (SVM) [12]. Additionally, a greedy forward feature selection method is employed, with the separability criterion based on the Mahalanobis distance (FSel-M) or the leave-one-out NN error (FSel-NN) [3].

Fig. 2 shows the average performance of the RNQC as a function of $|R|$. The error curves are compared to some variants of the NN rule. The 1NN-final and the kNN-final stand for the NN results obtained by using the entire training set T, hence such errors are plotted as horizontal lines. They are our reference. kNN is the k-NN rule directly applied to $D(T, R)$, while the kNN-DS is the Euclidean distance k-NN rule computed in $D(T, R)$ dissimilarity spaces (this means that a new Euclidean distance representation is derived from the vectors $D(x, R)$). In both cases, R is randomly selected. EdiCon-1NN presents the 1-NN result for the prototypes chosen by the editing and condensing criterion. The k-NN rule optimizes k in a LOO procedure over T.

In general, the best results are found for the feature selection approach with the separability criterion based on the Mahalanobis distance. In such a case, the RNQC in a dissimilarity space improves over the k-NN defined on all 523 training objects. This is achieved already for R consisting of $\sum_i \lceil \sqrt{|\omega_i|} \rceil \approx 44$ examples. Other selection methods need more prototypes.

Chicken data. The chicken pieces silhouettes set[2] consists of 446 binary images from chicken pieces: wing (117 examples), back (76), drumstick (96), thigh and

[2] `http://algoval.essex.ac.uk/data/sequence/chicken`

Fig. 3. Example chicken pieces: wing, back, drumstick, thigh-and-back and breast.

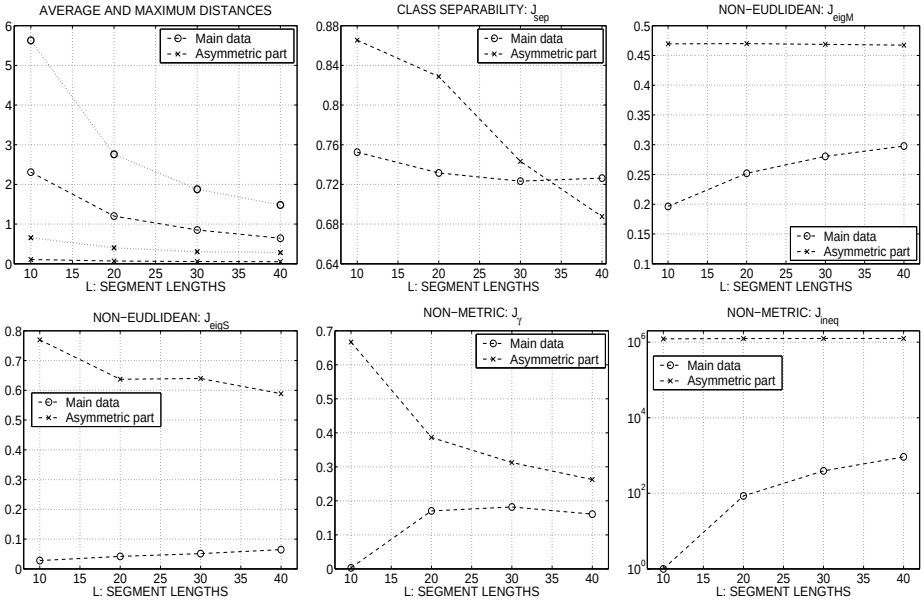

Fig. 4. Indices characterizing dissimilarity data.

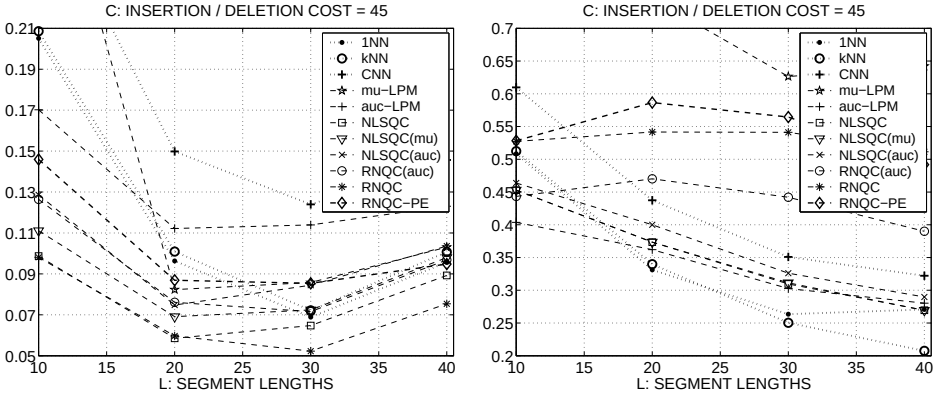

Fig. 5. Chicken data. Average (over 50 runs) 2-fold cross-validation errors for four values of L. Errors referring to the same classifier are connected by lines to enhance the visibility. The standard deviations of the average errors are 0.002 on average and the maximum of 0.003 for all classifiers.

back (61), and breast (96); see Fig. 3. The edges of the pieces are approximated by straight line segments of a fixed length L, $L = 10, 20, 30, 40$. Since the pieces are placed in arbitrary positions and mirror symmetry occurs, the initial string representation is by a sequence of angles between the neighboring segments. Then the edit distance is computed with fixed insertion and delation costs $C = 45$

(degrees) and a substation cost of the absolute difference between the angles; see [17]. Since the distances are asymmetric, we make them symmetric by averaging, i.e. $D^s = ((d_{ij}+d_{ji})/2)$. Additionally, we analyze the remaining asymmetric part by considering the (symmetric) representation $D^a = (|(d_{ij} - d_{ji})/2|)$.

The dissimilarity data are characterized by the indices described before, and illustrated in Fig. 4. We can observe that the average dissimilarities decrease with growing L. The smaller L, the larger maximal distances. None of the dissimilarity data set has a Euclidean behavior. Concerning the main data (D^s), the separability improves or remains constant, and the deviation from the Euclidean and non-metric behavior increases, both with the increase of L. Concerning D^a, with the increase of L, the separability improves, and the deviation from the Euclidean and non-metric behavior slightly decreases, but it is still huge.

In our study we perform 50 runs of 2-fold cross-validation on D^s and D^a, respectively, and average out the final results. In each cross-validation, all dissimilarities are additionally scaled by $\sqrt{|T|}$ to avoid too large values, and the errors are weighted by prior probabilities. The following classifiers are used: the 1-NN and k-NN rules directly applied to the dissimilarity complete representation $D^{a/s}(T,T)$ (k is optimized in LOO approach), edited-and-condensed nearest neighbor (CNN), μ-LPM ($\mu = \max\{0.01, 1.3{\cdot}\text{NN-LOO-err}\}$, auc-LPM (with the trade-off parameter set to 20) [16], NSQLC and RNQC with $\kappa = 0.05$. The later is used in both pseudo-Euclidean spaces and in dissimilarity spaces. Additionally, the NSQLC is trained on the representation sets determined by μ-LPM and auc-LPM, and denoted as NSQLC(μ) and NSQLC(auc). Remember that the LPMs and the NSQLC determine $R \subset T$ and that all multi-class linear classifiers are derived in the one-against-all strategy. More results can be seen in [17].

The classification results for D^s and D^a are shown in Fig. 5. We observe that the performances of all classifiers trained on main data (D^s) improve with the increasing value of L up to $L = 30$ and then decreases. The NLSQC performs the best or the second best, after the RNQC in dissimilarity spaces if $L \geq 30$. In total, however, nearly all objects are used for the representation. The RNQC needs all training objects. Concerning the remaining asymmetric contribution (D^a), the classifiers generally improve with growing L. The 1-NN and k-NN rules perform the best. The auc-LPM and the variants of the NLSQC are the next best. Although the results are not very good, they indicate that some some discriminative power is still present in the remaining (hence usually neglected from the analysis) asymmetric part of the dissimilarity measure. How to make use of this information remains open for future research.

5 Conclusions

Proximity representations offer a natural way for integrating qualitative knowledge with quantitative learning methodologies. Clustering techniques and statistical classifiers can be constructed in vector spaces that represent proximity

information. They are competitive to the nearest neighbor approaches, independently whether the measure is Euclidean or metric.

Acknowledgments. This work is supported by the Dutch Organization for Scientific Research (NWO). We thank Dr Volker Roth for the ProDom data, and Barbara Spillmann and Prof. Horst Bunke for the chicken dissimilarity data.

References

1. Vapnik, V.: Statistical Learning Theory. John Wiley & Sons, Inc. (1998)
2. Roth, V., Laub, J., Buhmann, J., Müller, K.R.: Going metric: Denoising pairwise data. In: Advances in Neural Information Processing Systems. (2003) 841–856
3. Pękalska, E., Duin, R., Paclík, P.: Prototype selection for dissimilarity-based classifiers. Pattern Recognition **39** (2006) 189–208
4. Dubuisson, M., Jain, A.: Modified Hausdorff distance for object matching. In: Int. Conference on Pattern Recognition. Volume 1. (1994) 566–568
5. Bunke, H., Sanfeliu, A., eds.: Syntactic and Structural Pattern Recognition Theory and Applications. World Scientific (1990)
6. Pękalska, E., Duin, R.: The Dissimilarity Representation for Pattern Recognition. Foundations and Applications. World Scientific, Singapore (2005)
7. Esposito, F., Malerba, D., Tamma, V., Bock, H., Lisi, F.: Similarity and Dissimilarity. In: Analysis of Symbolic Data. Springer-Verlag (2000)
8. Jacobs, D., Weinshall, D., Gdalyahu, Y.: Classification with Non-Metric Distances: Image Retrieval and Class Representation. TPAMI **22** (2000) 583–600
9. Veltkamp, R., Hagedoorn, M.: State-of-the-art in shape matching. In Lew, M., ed.: Principles of Visual Information Retrieval. Springer (2001) 87–119
10. Pękalska, E., Paclík, P., Duin, R.: A Generalized Kernel Approach to Dissimilarity Based Classification. J. of Machine Learning Research **2** (2002) 175–211
11. Pękalska, E., Duin, R., Günter, S., Bunke, H.: On not making dissimilarities Euclidean. In: Joint IAPR Int. Workshops on S+SSPR. (2004) 1145–1154
12. Haasdonk, B.: Feature space interpretation of SVMs with indefinite kernels. TPAMI **25** (2005) 482–492
13. Laub, J., Müller, K.R.: Feature discovery in non-metric pairwise data. Journal of Machine Learning Research (2004) 801–818
14. Goldfarb, L.: A new approach to pattern recognition. In Kanal, L., Rosenfeld, A., eds.: Progress in Pattern Recognition. Volume 2. Elsevier Science Publishers BV (1985) 241–402
15. Graepel, T., Herbrich, R., Schölkopf, B., Smola, A., Bartlett, P., Müller, K.R., Obermayer, K., Williamson, R.: Classification on proximity data with LP-machines. In: Int. Conference on Artificial Neural Networks. (1999) 304–309
16. Tax, D., Veenman, C.: Tuning the hyperparameter of an auc-optimized classifier. In: BNAIC. (2005) 224–231
17. Pękalska, E., Harol, A., Duin, R., Spillmann, B., Bunke, H.: Non-Euclidean or non-metric measures can be informative,. In: Joint IAPR Int. Workshops on S+SSPR. (2006) submitted
18. Borg, I., Groenen, P.: Modern Multidimensional Scaling. Springer-Verlag, New York (1997)
19. Ng, A., Jordan, M., Weiss, Y.: On spectral clustering: Analysis and an algorithm. In: Advances in Neural Information Processing Systems. Volume 14. (2002)
20. Corpet, F., Servant, F., Gouzy, J., Kahn, D.: Prodom and prodom-cg: tools for protein domain analysis and whole genome comparisons. Nucleid Acids Research **28** (2000) 267–269