On refining dissimilarity matrices for an improved NN learning

Robert P.W. Duin
ICT group, Faculty of EEMCS
Delft University of Technology, The Netherlands
r.duin@ieee.org

Elżbieta Pękalska
School of Computer Science
University of Manchester, United Kingdom
pekalska@cs.man.ac.uk

Abstract

Application-specific dissimilarity functions can be used for learning from a set of objects represented by pairwise dissimilarity matrices in this context. These dissimilarities may, however, suffer from various defects, e.g. when derived from a suboptimal optimization or by the use of non-metric or noisy measures. In this paper, we study procedures for refining such dissimilarities. These methods work in a representation space, either a dissimilarity space or a pseudo-Euclidean embedded space. On a series of experiments we show that refining may significantly improve the nearest neighbor classifications of dissimilarity measurements.

1. Introduction

Proximity measures have become popular in statistical learning as they naturally encode commonality between pairs of objects or groups of objects (e.g. clusters). Proximity measures can be used for deriving a numerical representation in which every element encodes similarity between pairs of objects. Such a measure is defined on raw or preprocessed measurements or task-specific features.

Learning from proximity data usually relies on either kernel methods for specifically designed kernels or on the nearest neighbor (NN) rule. Kernels are positive definite functions, interpreted as generalized inner products, hence similarity functions, in a Hilbert space induced by the kernel [15]. Kernel methods are powerful, but cannot handle arbitrary proximity data without necessary corrections. The NN rule can work well in such cases, but suffers from local decisions. In practice, many proximity measures used for matching and object comparison [2, 8] are neither positive definite nor proper distances in Hilbert spaces. As a result, other dissimilarity-based learning techniques have become important. This led to the development of indefinite kernel methods [10, 6, 11] and learning in embedded spaces [11].

Here we consider learning procedures applied to an n-element training set $\mathcal{X}$ for which all pairwise dissimilarities are given or can be computed by a known procedure. Such dissimilarities may result from suboptimal optimization (as in template matching) or by incorporation of invariance [8, 7]. The $n \times n$ dissimilarity matrix $D(\mathcal{X}, \mathcal{X})$ is non-Euclidean if it cannot be isometrically embedded into a Euclidean space. This often occurs when the measure is based on $\min$ or $\max$ operations. $D(\mathcal{X}, \mathcal{X})$ is non-metric if any metric requirement is disobeyed, e.g. symmetry or the triangle inequality.

The quality of a dissimilarity measure determines the speed of learning: the number of objects needed to reach a desired performance. We will discuss methods to refine a dissimilarity matrix for a given training set. These procedures are based on improving the internal consistency within the dissimilarity matrix. They thereby implicitly improve the dissimilarity measure. First, the objects are embedded in a vector space, next this space is transformed, and finally the dissimilarity matrix is reconstructed from the distances in the transformed space. We will study both unsupervised and supervised transformation procedures. Examples will be analyzed that either improve or deteriorate the NN classification based on the dissimilarity matrix.

2. Refining procedures

Fig. 1 illustrates that the dissimilarity matrix might be refined for a suboptimal dissimilarity measure. It shows a set of clusters in a Euclidean space with the dissimilarity that measures the distance between the two most neighboring points in the corresponding clusters (the single-linkage distance). Some are shown on the plot. Due to the local emphasis of the chosen dissimilarity measure, the resulting dissimilarity matrix does not properly encode the information on the relative positions of the cluster centers and their shapes. E.g. the

978-1-4244-2175-6/08/$25.00 ©2008 IEEE

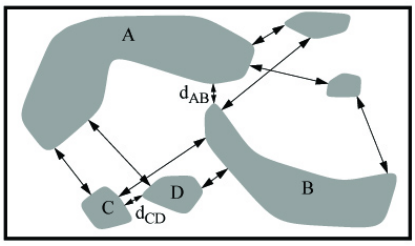

Figure 1. Single-linkage distance may be small for clusters which differ in position and shape.

clusters A and B have a small measured dissimilarity, but their centers and shapes are very different. This may be refined by considering the dissimilarity vectors representing dissimilarities from a particular cluster to all other ones. Such vectors differ for A and B, but are similar for C and D, which have a larger single-linkage distance, but are more similar in position and shape.

This intuitive example shows that a non-adequate dissimilarity measure may be improved by analyzing its behavior in *the context* of a set of objects. As a criterion for improvement we use the NN error on a test set in the final evaluation. Both supervised and unsupervised methods are considered for refining methods. A general procedure we use for this purpose is:

1) Given a dissimilarity matrix $D(\mathcal{X},\mathcal{X})$, map $\mathcal{X}$ into a representation (feature) vector space.
2) Consider a set of possible transformations there.
3) Optimize a chosen criterion in either a supervised or unsupervised way.
4) Reconstruct the dissimilarity matrix for the optimal transformation.
5) Evaluate the result.

Two representation spaces are considered in step 1), in which additional transformations will be considered. These are the pseudo-Euclidean embedded space (Section 2.1) and dissimilarity space (Section 2.2). As a criterion to optimize the transformations we use the leave-one-out NN error on the training set.

Re-scaling is used as a transformation of the representation space. First the dimensions are ranked according to some criterion, which is different for the two representation spaces we will discuss. Next, the scaling of every feature is optimized in a sequential procedure using the leave-one-out NN error of the training set as a criterion. This procedure is rather time consuming.

2.1. Pseudo-Euclidean embedding

A symmetric dissimilarity matrix $D := D(\mathcal{X},\mathcal{X})$ can be embedded in a pseudo-Euclidean space $\mathcal{E}$ by an isometric mapping [4, 11]. $\mathcal{E} = \mathbb{R}^{(p,q)} = \mathbb{R}^p \oplus \mathbb{R}^q$ is a vector space with a non-degenerate indefinite inner product $\langle\cdot,\cdot\rangle_{\mathcal{E}}$ such that $\langle\cdot,\cdot\rangle_{\mathcal{E}}$ is positive definite on $\mathbb{R}^p$ and negative definite on $\mathbb{R}^q$. So, we have $\langle \mathbf{x},\mathbf{y}\rangle_{\mathcal{E}} = \mathbf{x}^T \mathcal{J}_{pq}\mathbf{y}$, where $\mathcal{J}_{pq} = [I_{p\times p}\ 0;\ 0\ -I_{q\times q}]$ and I is the identity matrix. As a result, $d^2_{\mathcal{E}}(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T \mathcal{J}_{pq}(\mathbf{x}-\mathbf{y})$. The embedding relies on the indefinite Gram matrix G, derived as $G := -\frac{1}{2} J D^{\star 2} J$, where $D^{\star 2} = (d^2_{ij})$ and $J = I - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ is the centering matrix. The eigendecomposition of G leads to $G = Q\Lambda Q^T = Q|\Lambda|^{\frac{1}{2}}[\mathcal{J}_{pq};\ 0]|\Lambda|^{\frac{1}{2}}Q^T$, where Λ is a diagonal matrix of eigenvalues, first decreasing p positive ones, then increasing q negative ones, followed by zeros. Q is the matrix of eigenvectors. Since $G = X\mathcal{J}_{pq}X^T$ by definition of a Gram matrix, $X \in \mathbb{R}^n$ is found as $X = Q_n\,|\Lambda_n|^{\frac{1}{2}}$, where Q_n consists of n eigenvectors ranked according to their eigenvalues Λ_n. Note that X has a zero mean and is uncorrelated. The eigenvalues λ_i encode variances of the extracted features in $\mathbb{R}^{(p,q)}$.

The following distance measures between $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$ may be considered for refining the dissimilarity matrix:

$$\begin{aligned}\rho_{PES}(x,y) &= (\sum_{i=1}^{p}[x_i - y_i]^2 - \sum_{i=p+1}^{p+q}[x_i - y_i]^2)^{1/2} \\ &= (\sum_{i=1}^{n}\delta(i,p)[x_i - y_i]^2)^{1/2},\end{aligned}$$

where $\delta(i,p) = \text{sign}(p - i + 0.5)$. Since the complete pseudo-Euclidean embedding is perfect, $D(x,y) = \rho_{PES}(x,y)$ holds. Contributions from the q negative directions might be neglected (when assumed to reflect noise) and only the first p eigenvectors can be used:

$$\rho_{PES+}(x,y) = (\sum_{i=1}^{p}[x_i - y_i]^2)^{1/2} \qquad (1)$$

We may also compute distances in the associated Euclidean space by neglecting the minus-sign in $\mathcal{J}_{pq}$:

$$\rho_{AES}(x,y) = (\sum_{i=1}^{n}[x_i - y_i]^2)^{1/2} \qquad (2)$$

As also discussed, a transformation may be optimized for the training set by re-scaling eigenvectors. We therefore consider the following measure:

$$\rho_{\alpha\_PES}(x,y) = (\sum_{i=1}^{p}\alpha_i\delta(i,p)[x_i - y_i]^2)^{1/2} \qquad (3)$$

α_i is optimized by using the NN criterion over the training set. This is done in a sequential procedure applied to the eigenvectors sorted according to the decreasing magnitudes of the corresponding eigenvalues.

2.2. Dissimilarity space

Let $\mathcal{X} = \{x_1,\ldots,x_n\}$ be a training set. Given a dissimilarity function and/or dissimilarity data, we define a data-dependent mapping $D(\cdot,R) : \mathcal{X} \to \mathbb{R}^k$

from $\mathcal{X}$ to the so-called *dissimilarity space* [3, 5, 13]. The k-element set R consists of objects representative for the problem. This set is called the representation or prototype set and it may be a subset of $\mathcal{X}$. In the dissimilarity space each dimension $D(\cdot, p_i)$ describes a dissimilarity to a prototype p_i from R. In this paper, we initially choose $R := \mathcal{X}$. As a result, every object is described by an n-dimensional dissimilarity vector $D(x, \mathcal{X}) = [d(x, x_1) \ \ldots \ d(x, x_n)]^T$. The resulting vector space is endowed with the traditional inner product and the Euclidean metric.

Any dissimilarity measure ρ can be defined in this dissimilarity space. For the Euclidean distance, one has:

$$\rho_{DS}(x, y) = (\sum_{i=1}^{n} [d(x, x_i) - d(y, x_i)]^2)^{1/2} \quad (4)$$

In order to apply the earlier discussed re-scaling transformations the dimensions (features) of the dissimilarity space are ranked such that the sum of all distances to the corresponding objects is maximized for the next object to be selected.

$$\rho_{\alpha\_DS(x,y)} = (\sum_{i=1}^{n} \alpha_i [d(x, x_i) - d(y, x_i)]^2)^{1/2} \quad (5)$$

α_i is optimized by using the NN criterion over the training set in a sequential procedure over the dissimilarity features sorted as mentioned above.

The dissimilarity space is not directly affected by non-Euclidean characteristics of the data. The question is whether this still can make a difference. A simple correction is to add $2|\lambda_{\min}|$ to all off-diagonal elements of the squared dissimilarity matrix as $d_c(x_i, x_j) = (d^2(x_i, x_j) + 2|\lambda_{\min}|)^{1/2}$, $i \neq j$, where $\lambda_{\min}$ is the smallest negative eigenvalue found in the pseudo-Euclidean embedding. This results in a refined Euclidean-embeddable dissimilarity matrix [12]:

$$\rho_{cDS}(x, y) = (\sum_{i=1}^{n} [d_c(x, x_i) - d_c(y, x_i)]^2)^{1/2} \quad (6)$$

3. Experiments

The above refinement procedures are applied to a series of datasets. Due to space limits we mention just a few characteristics and a reference. Most of them are also used in [11].

- Chicken, dissimilarities based on the weighted edit distances between 446 shapes representing five classes of chicken pieces. They depend on two parameters [12]. We used Chicken_10_45, Chicken_29_45 and Chicken_40_45.
- Zongker, dissimilarities between 2000 handwritten digits in 10 classes based on deformable template matching [9]. We used a randomly selected subset of 400 digits, 40 out of every class.
- Polydistm57, modified Haussdorff distances between two classes of artificially generated polygons. A subset of 400 objects is used.
- WoodyPlants, dissimilarities between the shapes of 7634 leaves of 245 different species [1]. We used a subset of 400 leaves from 100 species.
- Cat-cortex, 65 objects in four classes represented by ordinal dissimilarity values [14].
- Newsgroups, 600 messages in four newsgroups related by a non-metric correlation measure [11].
- Protein, 213 protein sequences represented by the pairwise dissimilarities based on the concept of an evolutionary distance [5].
- Sonar, vectorial data based on 208 sonar signals represented by 60 features and two classes (http://www.ics.uci.edu/~mlearn/MLRepository.html). In the feature space we compute three distance matrices based on the $l1$-norm, $l2$-norm and on $D2$ - squared Euclidean distances.

All experiments are based on 2-fold cross-validation, repeated 5 times. Representation spaces are defined by all data examples. The training sets are used in these space to optimize the supervised refinement procedures, indicated by a '*' in Table 1. This table shows the average NN errors on the test objects for the reconstructed dissimilarity matrices. In brackets the standard deviations of the mean estimates are given. The first column (NEC) provides an index of non-Euclidean behavior of the data, derived from the pseudo-Euclidean embedding; $NEC = \sum_{j=p+1}^{p+q} |\lambda_j| / \sum_{i=1}^{p+q} |\lambda_i| \in [0, 1]$ and 0 indicates Euclidean distances. The second column is the error found for the original dissimilarity matrix. Refinements results that improve more than the sum of the standard deviations are underlined.

4. Discussion and Conclusion

As observed in Table 1, the NN performance can be improved for almost all non-vectorial dissimilarity measures. In these experiments, the results do not depend on the possible non-Euclidean characteristics of the data. The refinement of non-Euclidean dissimilarities defined by $l1$-norm or $D2$ on the vectorial Sonar data does not lead to NN improvements, while the almost Euclidean Protein data can be improved significantly. It is also interesting that the refinements based on either the straightforward neglection of negative directions in the pseudo-Euclidean embedding (ρ_{PES+}) or by using the associated Euclidean space (ρ_{AES}) hardly improve and sometimes even significantly deteriorate the result. It is encouraging that the unsupervised and thereby fast procedures in the dissimilarity space (ρ_{DS} and ρ_{cDS}) lead to very good results. The dissimilarity space seems

Table 1. Classification errors based on 5 times 2-fold cross-validation. Listed are the size, the NEC, the mean errors*1000 for the original data and for all refinement procedures (1)-(6) and their standard deviations*1000.

Data	#obj/class	NEC	Original	ρ_{PES+}	ρ_{AES}	ρ_{α_PES*}	ρ_{DS}	ρ_{α_DS*}	ρ_{cDS}
Cat-cortex	65 / 4	0.208	166(6)	74(13)	98(6)	139(13)	111(14)	98(14)	105(11)
News_groups	600 / 4	0.202	299(8)	324(7)	363(5)	301(8)	308(5)	299(6)	306(5)
Protein	213 / 4	0.001	32(5)	29(4)	24(4)	7(4)	6(1)	7(4)	5(3)
Sonar_I2	208 / 2	0.000	218(9)	218(9)	218(9)	225(9)	237(8)	224(9)	227(8)
Sonar_D2	208 / 2	0.288	218(9)	225(5)	223(3)	240(15)	250(8)	258(15)	248(12)
Sonar_I1	208 / 2	0.166	197(8)	200(8)	210(10)	208(6)	226(8)	228(9)	225(12)
Chicken_10_45	446 / 5	0.282	223(3)	417(3)	514(6)	141(5)	183(9)	174(8)	157(9)
Chicken_29_45	446 / 5	0.351	78(8)	185(8)	388(6)	62(7)	71(3)	64(4)	61(6)
Chicken_40_45	446 / 5	0.365	98(3)	223(7)	411(11)	97(4)	87(7)	80(4)	86(4)
Polydistm57	400 / 2	0.279	77(4)	75(6)	73(7)	64(11)	73(4)	71(2)	73(4)
Zongker	400 / 10	0.340	115(4)	139(6)	314(12)	69(6)	84(6)	85(5)	41(4)
WoodyPlants	400 / 100	0.192	501(9)	531(7)	583(9)	504(8)	643(7)	639(9)	583(11)

thereby a very good representation for refining dissimilarities by relating the dissimilarity vectors in the *context* of a large data set.

It has to be remarked that all our experiments include the test set in the definition of the representation spaces. This slows down the classification, but is an interesting procedure to make use of partially unlabeled data sets (semi-supervised learning).

In summary, we draw the following conclusion. Dissimilarity measurements can be improved in the *context of a larger data set*, but it is not yet clear when this is possible. Results have to be related to the characteristics of the dissimilarity measures that produce the original data. The next step is to describe the characteristics formally and derive the proper refinement procedures.

Acknowledgements. We thank colleagues for the data. See the references related to the data description.

References

[1] G. Agarwal, P. Belhumeur, S. Feiner, D. Jacobs, J. W. Kress, N. B. R. Ramamoorthi, N. Dixit, H. Ling, D. Mahajan, R. Russell, S. Shirdhonkar, K. Sunkavalli, and S. White. First steps toward an electronic field guide for plants. *Taxon*, 55:597–610, 2006.

[2] H. Bunke and A. Sanfeliu, editors. *Syntactic and Structural Pattern Recognition Theory and Applications*. World Scientific, 1990.

[3] R. Duin, D. de Ridder, and D. Tax. Experiments with object based discriminant functions; a featureless approach to pattern recognition. *Pattern Recognition Letters*, 18(11-13):1159–1166, 1997.

[4] L. Goldfarb. A new approach to pattern recognition. In L. Kanal and A. Rosenfeld, editors, *Progress in Pattern Recognition*, volume 2, pages 241–402. Elsevier Science Publishers BV, 1985.

[5] T. Graepel, R. Herbrich, P. Bollmann-Sdorra, and K. Obermayer. Classification on pairwise proximity data. In *Advances in Neural Information System Processing 11*, pages 438–444, 1999.

[6] B. Haasdonk. Feature space interpretation of SVMs with indefinite kernels. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 25(5):482–492, 2005.

[7] B. Haasdonk and H. Burkhardt. Invariant kernel functions for pattern analysis and machine learning. *Machine Learning*, 68(1):35–61, 2007.

[8] D. Jacobs, D. Weinshall, and Y. Gdalyahu. Classification with Non-Metric Distances: Image Retrieval and Class Representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(6):583–600, 2000.

[9] A. Jain and D. Zongker. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19(2):153–158, 1997.

[10] C. Ong, S. Mary, X.and Canu, and S. A.J. Learning with non-positive kernels. In *Int. Conference on Machine Learning*, pages 639–646, Brisbane, Australia, 2004.

[11] E. Pękalska and R. Duin. *The Dissimilarity Representation for Pattern Recognition. Foundations and Applications*. World Scientific, Singapore, 2005.

[12] E. Pękalska, A. Harol, R. Duin, B. Spillmann, and H. Bunke. Non-euclidean or non-metric measures can be informative. In *Joint IAPR Int. Workshops on SSPR*, pages 871–880, 2006.

[13] E. Pękalska, P. Paclík, and R. Duin. A Generalized Kernel Approach to Dissimilarity Based Classification. *J. of Machine Learning Research*, 2(2):175–211, 2002.

[14] J. Scannell, C. Blakemore, and M. Young. Analysis of connectivity in the cat cerebral cortex. *Journal of Neuroscience*, 15:1463–1483, 1995.

[15] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.