Journal of Biomedical Informatics 44 (2011) 775–788

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Class proximity measures – Dissimilarity-based classification and display of high-dimensional data

R.L. Somorjai [a,*], B. Dolenko [a], A. Nikulin [a], W. Roberson [a], N. Thiessen [b]

[a] Institute for Biodiagnostics, National Research Council Canada, 435 Ellice Avenue, Winnipeg, MB, Canada R3B 1Y6
[b] Genome Sciences Centre, BC Cancer Agency, 570 West 7th Ave. – Suite 100, Vancouver, BC, Canada V5Z 4S6

ARTICLE INFO

Article history:
Received 18 October 2010
Available online 27 April 2011

Keywords:
Mappings
Projections
Class-proximity planes
High-dimensional data
Proximity measures
Distance/dissimilarity measures
Visualization
Classification

ABSTRACT

For two-class problems, we introduce and construct mappings of high-dimensional instances into dissimilarity (distance)-based Class-Proximity Planes. The Class Proximity Projections are extensions of our earlier relative distance plane mapping, and thus provide a more general and unified approach to the simultaneous classification and visualization of many-feature datasets. The mappings display all L-dimensional instances in two-dimensional coordinate systems, whose two axes represent the two distances of the instances to various pre-defined proximity measures of the two classes. The Class Proximity mappings provide a variety of different perspectives of the dataset to be classified and visualized. We report and compare the classification and visualization results obtained with various Class Proximity Projections and their combinations on four datasets from the UCI data base, as well as on a particular high-dimensional biomedical dataset.

1. Introduction

The promise and potential of noninvasive diagnosis/prognosis of diseases and disease states is the principal reason for acquiring specific types of biomedical data (e.g., spectra or gene microarrays) from biofluids and tissues. However, such data are characterized by relatively few *instances* (samples) (N = O(10)–O(100)), initially in a very *high-dimensional feature space*, with dimensionality L = O(1000)–O(10,000). These two characteristics lead to the twin curses of *dimensionality* and *dataset sparsity* [38]. Any analysis of such data calls for special considerations. To lift the curse of dimensionality, the obvious, standard course of action is to carry out feature selection/extraction/generation (this may be avoided by using kernel SVMs, but with a concomitant loss of interpretability). Many approaches are possible. A particular, powerful version is a dissimilarity (distance)-based, dimensionality-reducing mapping/projection from L to two dimensions (only two, because we shall consider only 2-class problems. Higher dimensional mappings are also feasible; however, in more than three dimensions the classification results cannot be visualized.) Naturally, mapping to lower dimensions inevitably leads to information loss; hence, not all original distances can be preserved *exactly*. A number of projection methods, e.g., Isomap [42], Multidimensional Scaling [6], etc. attempt to minimize the projection errors for *all* distances. However, as we have shown earlier for the relative distance plane (RDP) mapping [39], the *exact* preservation of all distances is not necessary for an *exact visualization* in some *distance plane*. The relative distance plane is created by any instance's two relative distances (two *new features*) to any pair of other instances, one from each class. Furthermore, the mapping appears to preserve whatever *class separation* exists in the original feature space. A *direct classification* in this projected distance plane is then feasible and may reveal additional, useful information about the originally high-dimensional dataset (see e.g., [41]).

Encouraged by the successes and promise of the RDP mapping, in this work we explore and extend this dissimilarity-based concept. Of course, dissimilarity/distance-based classification is not new. In fact, all nearest neighbor classifiers [9,11] are distance based. "Instance-based classification" is a generalization of this model-free approach, also using nearest neighbor concepts (e.g., [44]). Since the early 2000s, Duin's group has been advocating distance-based classification, e.g., in [33,35,29,30,34], etc. For a thorough discussion of many relevant theoretical and practical issues, see [36].

The possibility of converting L-dimensional (L arbitrary) datasets into 2-dimensional equivalents via general and adaptable distance-based mappings is very attractive and has important implications, especially for $L \gg N$. The most general conceptual extension and subsequent implementation of such a mapping require selecting both a *class proximity measure* π and a *distance/dissimilarity measure* Δ. (Note that dissimilarity is the more general concept and does not have to be a metric.) We may create a flexible

* Corresponding author. Fax: +1 204 984 5472.
E-mail address: Ray.Somorjai@nrc-cnrc.gc.ca (R.L. Somorjai).

1532-0464/$ - see front matter Crown Copyright © 2011 Published by Elsevier Inc. All rights reserved.
doi:10.1016/j.jbi.2011.04.004

framework for both classification and visualization by generating a wide range of $[\pi; \Delta]$ pairs. An important extension of this mapping process is the use of *class-dependent* $[\pi; \Delta]$ pairs, i.e., in its most general form $[\pi_1^\lambda; \Delta_1^\lambda]$ for class 1, $[\pi_2^\nu; \Delta_2^\nu]$ for class 2. An example is *Regularized Discriminant Analysis*, with different Δ_s^κ for the two classes [18,15]. *Quadratic Discriminant Analysis* is a special case. Other $[\pi^\lambda; \Delta^\lambda]$ choices are (K_1, K_2)-NN classifiers, $K_1 \neq K_2$ [19]. Paredes and Vidal introduced a class-dependent dissimilarity measure [31]. Class-dependent PCA/PLS is often used in the popular software SIMCA-P ([12]).

Here we introduce and discuss a few of the many possibilities, both for distance/dissimilarity and class proximity measures. We implemented the majority of these in our software CPP (Class Proximity Projector, *vide infra*). The major goal and thrust of the paper is threefold: (a) Choose different class proximity measures (π^vs) to *represent* the two classes in different ways. (b) Select a distance measure Δ. (c) With this Δ, compute and display the two distances $d_1(\mathbf{Z})$, $d_2(\mathbf{Z})$ of an instance $\mathbf{Z}$ to the chosen π^v.

The article is organized as follows. In the Introduction, we already discussed the motivation for using a distance-dependent approach for both the visualization and classification of high-dimensional data. Next, we define and list the various common distance measures we may use. This is followed by the description of several possible class representations and class proximity measures. In particular, we introduce, discuss and compare four major categories for representing and positioning an instance in a Class Proximity (CP) plane. In the Results and Discussion section we first illustrate in detail, on a high-dimensional biomedical (metabolomic) dataset (^{1}H NMR spectra of a biofluid) several feasible possibilities and processes, based on concepts of the Class Proximity Projection approach. We repeat this process for four datasets from the UCI Repository. We conclude with general observations and a summary.

1.1. Distance measures Δ

Consider a two-class dataset, and assume that the data comprise N instances in an L-dimensional feature space. Thus the original mth instance vector $\mathbf{Z}^{(m)}$ has L components $[Z_1^{(m)}, \ldots, Z_L^{(m)}]$, $m = 1, \ldots, N$, $N = N_1 + N_2$, with N_k instances in class k, $k = 1, 2$.

For computing the distance δ_{mn} between instances $\mathbf{Z}^{(m)}$ and $\mathbf{Z}^{(n)}$, we have implemented several distance measures Δ (throughout, the superscript t denotes the *transpose*).

1.1.1. Minkowski (MNK) distance

$$\delta_{mn}(\mathrm{MNK}(\gamma)) \equiv \delta(\mathbf{Z}^{(m)},\ \mathbf{Z}^{(n)};\quad \mathrm{MNK}(\gamma)) = \{\Sigma_{k=1}^{L} \mid Z_k^{(m)} - Z_k^{(n)}|^\gamma\}^{1/\gamma}$$

$\gamma = 1, 2, \infty$ correspond to the Manhattan, Euclidean and Chebychev (max norm) distances, respectively; γ is an optimizable parameter.

The following distance measures require the class covariance matrices $\mathbf{S}_1$ and $\mathbf{S}_2$. All analytical formulae presented assume that the class distributions are Gaussian; however, we shall also use them for arbitrary distributions.

1.1.2. Anderson–Bahadur (AB) distance

$$\begin{aligned}\delta_{mn}(\mathrm{AB}(\alpha)) &\equiv \delta(\mathbf{Z}^{(m)},\ \mathbf{Z}^{(n)};\quad \mathrm{AB}(\alpha)) \\ &= [(\mathbf{Z}^{(m)} - \mathbf{Z}^{(n)})^t \mathbf{S}(\alpha)^{-1} (\mathbf{Z}^{(m)} - \mathbf{Z}^{(n)})]^{1/2}\end{aligned}$$

$\mathbf{S}_1$, $\mathbf{S}_2$ are the training set estimates of the class covariance matrices.

$$\mathbf{S}(\alpha) = (1-\alpha)\mathbf{S}_1 + \alpha\mathbf{S}_2;\quad 0 \le \alpha \le 1$$

The optimizable parameter α controls the amount of mixing between $\mathbf{S}_1$ and $\mathbf{S}_2$; when $\alpha = 0.5$, we obtain the *Mahalanobis* distance and $\mathbf{S}(0.5)$ is the pooled covariance matrix used in Fisher's linear discriminant analysis (FLD). [1].

1.1.3. Symmetric Kullback–Leibler (SKL) "distance"

$$\begin{aligned}\delta_{mn}(\mathrm{SKL}) &\equiv \delta(\mathbf{Z}^{(m)},\ \mathbf{Z}^{(n)};\quad \mathrm{SKL}) \\ &= [\{(\mathbf{Z}^{(m)} - \mathbf{Z}^{(n)})^t (\mathbf{S}_1^{-1} + \mathbf{S}_2^{-1})(\mathbf{Z}^{(m)} - \mathbf{Z}^{(n)}) + tr[\mathbf{S}_1^{-1}\mathbf{S}_2 \\ &\quad + \mathbf{S}_2^{-1}\mathbf{S}_1 - 2\mathbf{I}_M]\}/2]^{1/2}\end{aligned}$$

where $\mathbf{I}_M$ is the M-dimensional unit matrix. When $\mathbf{S}_1 = \mathbf{S}_2$, $\delta_{mn}(\mathrm{SKL})$ is proportional to the Mahalanobis distance. $\delta_{mm}(\mathrm{SKL}) \neq 0$ unless $\mathbf{S}_1 = \mathbf{S}_2$, hence $\delta_{mn}(\mathrm{SKL})$ is not truly a distance.

1.1.4. Cosine "distance"

$$\begin{aligned}\delta_{mn}(\mathrm{COS}) &\equiv \delta(\mathbf{Z}^{(m)},\ \mathbf{Z}^{(n)};\quad \mathrm{COS}) = 1 - \mathbf{Z}^{(m)} \cdot \mathbf{Z}^{(n)} / (\|\mathbf{Z}^{(m)}\| \|\mathbf{Z}^{(n)}\|) \\ &= 1 - \cos\theta\end{aligned}$$

where θ is the angle between the vectors $\mathbf{Z}^{(m)}$ and $\mathbf{Z}^{(n)}$.

1.2. Class proximity measures

The proposed *projection/mapping* procedure requires and uses a proximity/distance measure pair $[\pi; \Delta]$. More specifically, we write $[\pi_k; \Delta_k]$, $k = 1, 2$, when we want to refer explicitly to the specific class representation. For any multivariate instance vector $\mathbf{Z}^{(m)}$, whatever its class, we compute the two projected coordinates $Y[\mathbf{Z}^{(m)}] = d_1(\mathbf{Z}^{(m)}|\pi_1;\ \Delta_1)$, $X[\mathbf{Z}^{(m)}] = d_2(\mathbf{Z}^{(m)}|\pi_2;\ \Delta_2)$ with respect to the chosen $[\pi; \Delta]$; the $d_k(\mathbf{Z}^{(m)}|\pi;\ \Delta)$ are the $[\pi; \Delta]$-generated distances of $\mathbf{Z}^{(m)}$ to class k, $k = 1, 2$. We may readily display this two-dimensional representation $(Y[\mathbf{Z}^{(m)}],\ X[\mathbf{Z}^{(m)}])$ of $\mathbf{Z}^{(m)}$ in the appropriate $[\pi; \Delta]$ class-proximity plane (CP-Plane), thus allowing the visualization of the L-dimensional data, while essentially preserving class separation. In addition, direct classification in the CP-Plane will be possible [41]. Because the CP mappings are distance-based, we only need a single computation of a distance matrix $\mathbf{D} = [\delta_{mn}]$, $m, n = 1, \ldots, N$, where δ_{mn} is the distance between L-dimensional instances $\mathbf{Z}^{(m)}$ and $\mathbf{Z}^{(n)}$, calculated with the chosen distance measure Δ.

It is more general and frequently more advantageous to characterize the class representations in terms of *prototypes*, say $\mathbf{R}$. Prototype generation may range from maintaining the *status quo* (i.e., the prototypes are the N individual, original instances), to the other limiting case, one prototype per class, e.g., the two class centroids (the basis for the Nearest Mean Classifier, [21]). For intermediate cases, when there is more than one prototype per class, a number of possibilities exist. An excellent earlier review is [5]. Various instance reduction techniques to produce prototypes are discussed in [45]. Depending on ultimate requirements, different approaches were proposed in [32] and in [22]. Another attractive option is to carry out some version of class-dependent agglomerative clustering (e.g., k-means). The inputs are the distance measure Δ and the number of clusters R_r required for class r, $r = 1, 2$. The clustering algorithm partitions the two classes and redistributes the N_r instances in class r into m_{rc} instances in cluster c_r. The R_r cluster centroids, $r = 1, 2$, provide the $R_1 + R_2$ prototypes. Then, for any instance, some function of its distances to these prototypes (or to individual members of the clusters) provides the next stage for defining new class representations. From these distance functions, different class proximity measures may be generated. Amongst the more sophisticated prototype generation approaches, the instance-adaptive condensation schemes introduced and explored in [24] are noteworthy.

For any distance/dissimilarity matrix $\mathbf{D}$, the subscript of the class proximity measure π_k reflects the prototype-based *explicit* representation of class k. (As a specific example, if the proximity