2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

Simultaneous Visualization of Samples, Features and Multi-Labels

Mineichi Kudo and Keigo Kimura
Division of Computer Science
and Information Technology
Graduate School of Information
Science and Technology
Hokkaido University, JAPAN
E-mail: {mine,kkimura}@main.ist.hokudai.ac.jp

Michal Haindl
Department of Pattern Recognition
Institute of Information Theory
and Automation
Czech Academy of Sciences
Czech Republic
E-mail:haindl@utia.cas.cz

Hiroshi Tenmoto
Department of Information Engineering
National Institute of Technology
Kushiro College
Otanoshike Nishi 2-32-1, Kushiro
Hokkaido 084-0916, JAPAN
E-mail: tenmo@kushiro-ct.ac.jp

***Abstract*—Visualization helps us to understand single-label and multi-label classification problems. In this paper, we show several standard techniques for simultaneous visualization of samples, features and multi-classes on the basis of linear regression and matrix factorization. The experiment with two real-life multi-label datasets showed that such techniques are effective to know how labels are correlated to each other and how features are related to labels in a given multi-label classification problem.**

I. Introduction and Related works

Dimension reduction including visualization, as a special case, is important in processing time reduction, precision improvement, and problem understanding. The algorithms are divided into two categories of unsupervised approaches such as PCA and ICA, and supervised approaches such as Fisher LDA (for example, see [1]). Recently, in the viewpoint of sparse coding and matrix factorization, some novel dimension reduction techniques have been studied such as LLE and S-Isomap [2], [3], [4], [5], [6]. It also turns out that they are closely related to traditional statistical techniques above and to each other as shown in [7]. In this paper, therefore, returning to the original standing point, we discuss what information is available for dimension reduction, and then propose several fundamental ways for dimension reduction.

We concentrate on visualization with dimension two in this paper. In multi-label classification, it is important to know how single labels are related to multi-labels in addition to how samples are related to those labels. Visualization is one of strong tools for these goals. Nevertheless, such studies are a very few, e.g., [4], We propose three algorithms for simultaneous visualization of samples and multi-labels, and one more algorithm for simultaneous visualization of features and labels. Some of these algorithms are based on the relationship between regression with sparsity and low-rank matrix factorization. The sparsity is one of key aspects in regression, as seen in ridge regression and lasso [1], and is useful for dimension reduction. The matrix factorization with a low-rank constraint also contributes for dimension reduction as seen in [5], [6].

II. Fundamental Relationship

A. Notation

Let us firstly summarize necessary notations and basic relationships.

We consider n data $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ with class multi-labels $y_1, y_2, \ldots, y_n \in \{0,1\}^\ell$. Their matrix expression is given by

$$X = (x_1 \quad x_2 \quad \cdots \quad x_n) = \begin{pmatrix} x_{11} & \cdots & x_{n1} \\ x_{12} & \cdots & x_{n2} \\ \vdots & \ddots & \vdots \\ x_{1d} & \cdots & x_{nd} \end{pmatrix} \in \mathbb{R}^{d\times n}$$

$$Y = (y_1 \quad y_2 \quad \cdots \quad y_n) = \begin{pmatrix} y_{11} & \cdots & y_{n1} \\ \vdots & \ldots & \vdots \\ y_{1\ell} & \cdots & y_{n\ell} \end{pmatrix} \in \{0,1\}^{\ell\times n}$$

$$Z = (z_1 \quad z_2 \quad \cdots \quad z_n) = \begin{pmatrix} X \\ Y \end{pmatrix} \in \{\mathbb{R}\cup\{0,1\}\}^{(d+\ell)\times n}$$

We use superscript T for denoting the transpose of a matrix.

B. Regression and Matrix Approximation

We assume that X is already *centered* and *sphered* [1] such as

$$\mu_x = \frac{1}{n}X\mathbf{1} = \mathbf{0}, \quad \Sigma_x = \frac{1}{n}XX^T = I_d,$$

where $\mathbf{1}$ and $\mathbf{0}$ are the column vectors of all one's and all zero's of an appropriate dimension, respectively, and I_d is the identity matrix of dimension d. In classification, such a standardization is essential and necessary because the problems should be invariant for any affine transformation.

Especially, we notice the importance of sphering from the following relationship between a linear regression of Y on X and a matrix approximation of XY^T/n.

Theorem 1. *If $X \in \mathbb{R}^{d\times n}$ is sphered, then the linear regression of $Y \in \mathbb{R}^{\ell\times n}$ on X is equivalent to a matrix approximation problem of XY^T/n in the sense of*

$$\arg\min_{A\in\mathbb{R}^{d\times\ell}} \|Y - A^TX\|_F^2 = \arg\min_{A\in\mathbb{R}^{d\times\ell}} \|\frac{1}{n}XY^T - A\|_F^2, \quad (1)$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix.

[1] Centering is made by $X \leftarrow X(I - \frac{1}{n}\mathbf{1}\mathbf{1}^T)$ and, sphering is made by $X \leftarrow \Lambda^{-1/2}U^TX$ using the spectral decomposition $XX^T/n = U\Lambda U^T$.

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

Proof: This equation is derived as follows.

$$\begin{aligned} & \|Y - A^T X\|_F^2 \\ = \; & \mathrm{Tr}(Y - A^T X)^T (Y - A^T X) \\ = \; & \mathrm{Tr}(Y^T Y - 2Y^T A^T X + X^T A A^T X) \\ = \; & \mathrm{Tr}(Y^T Y - 2A^T X Y^T + A^T X X^T A) \\ = \; & \mathrm{Tr}(Y^T Y - 2A^T X Y^T + nA^T A) \quad (XX^T = nI_d) \\ = \; & c - 2\,\mathrm{Tr}(A^T X Y^T) + n\|A\|_F^2 \quad (c = \|Y\|_F^2), \end{aligned}$$

while

$$\begin{aligned} \|\frac{1}{n} XY^T \; - \; & A\|_F^2 \\ = \; & \mathrm{Tr}(\frac{1}{n} XY^T - A)^T (\frac{1}{n} XY^T - A) \\ = \; & \mathrm{Tr}(\frac{1}{n^2} Y X^T X Y^T - \frac{2}{n} A^T X Y^T + A^T A) \\ = \; & c' + \frac{1}{n}(-2\,\mathrm{Tr}(A^T X Y^T) + n\|A\|_F^2) \\ & (c' = \frac{1}{n^2}\,\mathrm{Tr}(Y X^T X Y^T)). \end{aligned}$$

This comparison completes the proof. ∎

Note that this proof shows a monotonic relationship between these two terms in addition to the fact that the same A minimizes them at the same time.

This theorem is important in the following sense: 1) the regression (LHS) gives a direct way of predicting $\hat{y}$ from x, 2) the approximation (RHS) shows how we can approximate the relationship between features and labels (classes), and 3) this theorem bridges above two different concepts and make it possible to find the best low-rank representation as will be shown below. Without such a low-rank requirement, the solution is obviously $A = (XY^T)/n$ from RHS of (1), thus, LHS of (1) becomes $\|Y - A^T X\|_F^2 = \|Y - (1/n) Y X^T X\|_F^2$.

On the other hand, the importance of centering is shown as

Corollary 2. *If $X \in \mathbb{R}^{d\times n}$ is sphered and centered, then the solution A for RHS of (1) is invariant to any shift of y, that is, the solution A is the same even for $Y - y_0 \mathbf{1}^T$ with any $y_0 \in \mathbb{R}^\ell$. While, LHS of (1) is minimized at $y_0 = \mu_0$. Thus, with centered Y, $\bar{Y} = Y - \mu_y \mathbf{1}^T$, we have*

$$\arg\min_{A\in\mathbb{R}^{d\times\ell}} \|\bar{Y} - A^T X\|_F^2 = \arg\min_{A\in\mathbb{R}^{d\times\ell}} \|\frac{1}{n} XY^T - A\|_F^2. \quad (2)$$

Proof: The shift invariance property is shown as, for any y_0,

$$\begin{aligned} & \|\frac{1}{n} X(Y - y_0\mathbf{1}^T)^T - A\|_F^2 \\ = \; & \|\frac{1}{n} XY^T - \frac{1}{n} X\mathbf{1} y_0^T - A\|_F^2 \\ = \; & \|\frac{1}{n} XY^T - A\|_F^2 \quad (\mu_x = (1/n) X\mathbf{1} = \mathbf{0}) \end{aligned}$$

This means that if Y changes to $Y - y_0\mathbf{1}^T$, only the norm in the LHS changes. It is easy to show that the norm is minimized with $y_0 = \mu_y$, which is the best for regression. From Theorem 1 and $y_0 = \mu_y$, (2) is obtained. ∎

In the viewpoint of regression, we should add an extra dimension with value one to each x_i for allowing the regression to have a constant term. However, from Corollary 2, we see that the same thing is possible by $\hat{y} = A^T x + \mu_y$ with the solution A of (2).

C. Low-Rank Approximation

In this study, we are interested in when A has a low rank k ($\leq \min\{d, \ell\}$). Under this *low-rank* requirement, the approximation problem has a clear solution. From Schmidt's theorem [10] (Chap. 6.3), when $(1/n)XY^T$ is singular value decomposed into $U\Lambda V^T$, the solution is given by

$$A = U_{(k)} \Lambda_{(k)} V_{(k)}^T, \quad \text{s.t.} \quad U_{(k)}^T U_{(k)} = V_{(k)}^T V_{(k)} = I_k, \quad (3)$$

where the matrices with suffix (k) correspond to those of the largest k eigenvalues. With this solution, the regression criterion gives us a way of prediction:

$$\begin{aligned} \dot{y} \; & = \; A^T x = V_{(k)} \Lambda_{(k)} U_{(k)}^T x & (4) \\ \hat{y} \; & = \; \mathrm{Bin}(\dot{y} + \mu_y). & (5) \end{aligned}$$

where operation $\mathrm{Bin}(\cdot)$ binalizes each component value to zero or one by threshold 1/2 and μ_y is added from Corollary 2.

D. Training Information

Let us examine how many different ways are possible in dealing with training information. Basically there are three ways as follows.

From the combined training information $Z^T = (X^T \;\; Y^T) \in \{\mathbb{R} \cup \{0,1\}\}^{n\times(d+\ell)}$, we can consider the first two of the three ways as

$$\begin{aligned} ZZ^T \; & = \; \begin{pmatrix} XX^T & XY^T \\ YX^T & YY^T \end{pmatrix} \in \mathbb{R}^{(d+\ell)\times(d+\ell)} & (6) \\ Z^T Z \; & = \; X^T X + Y^T Y \in \mathbb{R}^{n\times n}, & (7) \end{aligned}$$

When X and Y are both centered, the first formula (6) is essentially equivalent to the covariance matrix $\Sigma_Z = ZZ^T/n$. The third way is to use the connection between individual sample and its labels that is just given by

$$Y \in \{0,1\}^{n\times\ell}. \quad (8)$$

III. Visualization Techniques

On the basis of low-rank solution (3) and three kinds of information (6), (7) and (8), we propose several simultaneous visualization techniques.

A. Sample-Label Visualization

For visualization ($k = 2, 3$), the solution (3) gives a way to map the samples and labels into a same low-dimensional (k-dimensional) Euclidean space $\mathbb{R}^k$ as

(Visualization `SL-A`)

$$\begin{aligned} \tilde{x} \; & = \; \Lambda_{(k)} U_{(k)}^T x \in \mathbb{R}^k, \quad \tilde{y} = V_{(k)}^T (y - \mu_y) \in \mathbb{R}^k, \\ & \quad \text{where } \frac{1}{n} XY^T = U\Lambda V^T. \end{aligned}$$

Note that in single-label cases

$$\frac{1}{n}XY^T = (\frac{n_1}{n}\mu_1, \frac{n_2}{n}\mu_2, \ldots, \frac{n_\ell}{n}\mu_\ell), \tag{9}$$

where n_i is the appearance number of labels i and μ_i is the mean of samples of label i. Thus, the SVD of XY^T/n is that of the class means weighted by the estimated priors. Therefore, rank$(XY^T/n) \leq \ell - 1$. For multi-label classification, the rank is less than the number ℓ' of distinct label subsets seen in Y.

(An example) In two-class cases ($\ell = 2$) with centered X, since $n_2\mu_2 = -n_1\mu_1$, we have

$$\frac{1}{n}XY^T = \frac{1}{n}(n_1\mu_1, -n_1\mu_1) = \frac{n_1}{n}\mu_1(1, -1) \in \mathbb{R}^{d\times 2}.$$

Hence, $\frac{1}{n}XY^T$ has rank 1 and is decomposed into

$$U_{(1)}\Lambda_{(1)}V_{(1)}^T = \left(\frac{1}{\|\mu_1\|}\mu_1\right)\left(\frac{n_1}{n}\sqrt{2}\|\mu_1\|\right)\left(\frac{1}{\sqrt{2}} \ -\frac{1}{\sqrt{2}}\right).$$

The simultaneous low-dimensional mappings with $k = 1$ are given by

$$\tilde{x} = \Lambda_{(1)}U_{(1)}^T x = \frac{n_1}{n}\sqrt{2}\|\mu_1\|\frac{1}{\|\mu_1\|}\mu_1^T x = \frac{n_1}{n}\sqrt{2}\mu_1^T x$$

and

$$\tilde{y} = V_{(1)}^T(y - \mu_y) = \begin{cases} \sqrt{2}(1 - n_1/n) & (y = (1, 0)^T) \\ -\sqrt{2}n_1/n & (y = (0, 1)^T) \end{cases}.$$

The prediction $\hat{y}$ is, from (4) and (5), given by

$$\hat{y} \ = \ \dot{y} + \mu_y = \begin{pmatrix} n_1/n \\ -n_1/n \end{pmatrix}\mu_1^T x + \begin{pmatrix} n_1/n \\ 1 - n_1/n \end{pmatrix}.$$

B. Sample-Label Visualization Using Laplacian Eigenmap

According to the way of Laplacian eigenmap [9], we can have another technique for simultaneous low-dimensional mapping of samples and labels. Suppose that we are given three kinds of similarity at once: sample-sample similarity w_{ij}, label-label similarity w_{mo} and sample-label similarity w_{im}. Then the k-dimensional representation g_i of sample i and the k-dimensional representation h_m of label m are given by minimizing, under some normalization constraint,

$$\sum_{ij} w_{ij}\|g_i - g_j\|^2 + \sum_{mo} w_{mo}\|h_m - h_o\|^2 + 2\sum_{im} w_{im}\|g_i - h_m\|^2,$$

where i and j run over $\{1, 2, \ldots, n\}$, m and o run over $\{1, 2, \ldots, \ell\}$, and $g_i, h_m \in \mathbb{R}^k$. From the well-known equivalence of this sum and the objective function of Graph Laplacian, we can solve this problem by finding $U = ((g_i)\ (h_m)) \in \mathbb{R}^{k\times(n+\ell)}$ minimizing

$$\mathrm{Tr}(U(D - W)U^T) \quad \text{subject to} \quad UDU^T = I_k, \tag{10}$$

for

$$W = \begin{pmatrix} (w_{ij}) & (w_{im}) \\ (w_{mi}) & (w_{mo}) \end{pmatrix} \in \mathbb{R}^{(n+\ell)\times(n+\ell)}$$

and

$$D = \mathrm{diag}(\sum_j w_{1j}, \ldots, \sum_j w_{(n+\ell)j}).$$

In this paper, from (7) and (8), we propose their realization as

$$\begin{aligned} (w_{ij}) &= (X^TX + Y^TY)_{ij} \in \mathbb{R}^{n\times n}, \\ (w_{mo}) &= (YY^T)_{mo} \in \mathbb{R}^{\ell\times\ell}, \quad \text{and} \\ (w_{im}) &= (Y)_{im} \in \{0, 1\}^{n\times\ell}. \end{aligned}$$

Solving (10) as a generalized eigenvalue problem, we have the next visualization. Here, since one is the trivial largest eigenvalue, we use the largest k eigenvalues less than one.

(Visualization SL−B)

$$\tilde{x}_i = u_i \in \mathbb{R}^k, \quad \tilde{y}_m = u_{n+m} \in \mathbb{R}^k,$$

where, u_i is ith column of $U_{(k)} \in \mathbb{R}^{k\times(n+\ell)}$ satisfying

$$D^{-1}WU_{(k)}^T = U_{(k)}^T\Lambda_{(k)} \text{ for } W = \begin{pmatrix} X^TX + Y^TY & Y^T \\ Y & YY^T \end{pmatrix}.$$

It should be noted that, unlike SL−A method, this mapping is nonlinear and the transformation from a feature vector or a label vector to the corresponding points are not functionally realized. Therefore, we cannot map a newly arrived sample or label vector without a special treatment. Such a treatment is seen in the authors' another study [4]. In [4], to avoid the duplicating usage of Y^TY and YY^T, we proposed to use

$$W = \begin{pmatrix} \alpha X^TX & Y^T \\ Y & \beta YY^T \end{pmatrix}, \alpha, \beta > 0.$$

We also introduced locality as seen in LLE [2]: one binary relation is calculated from the neighborhood relationship between two samples, another binary relationship from the neighborhood relationship between two labels. For the mapping of an unseen sample, we proposed one linear mapping and one nonlinear mapping [4].

One problem of this approach is the common normalization applied for different kinds of source. The solution $U_{(k)}$ of this optimization is given by the largest k eigenvectors of $D^{-1}W$. The multiplication of D^{-1} affects row by row because D is diagonal. This means that the first n elements and the following ℓ elements in a row of W are divided by a same element of D. This is not always appropriate. Another problem is the unbalanced sizes of X^TX and YY^T. As the sample number n increases, the former size of X^TX increases too, but the latter size of YY^T stays as a constant. Accordingly, the effect of label-label relationship vanishes and the sample-label connection Y is weaken as well. As a result, the sample similarity, thus the positions of g_i's, dominate this optimization.

C. Feature-Label Visualization

In a similar way, we can have another technique for simultaneous low-dimensional representations of features and labels. We consider (6) and assume that X and Y are both centered and "norm-to-one" standardized ($X\mathbf{1} = \mathbf{0}, Y\mathbf{1} = \mathbf{0}, \mathrm{diag}(XX^T) = I_d, \mathrm{diag}(YY^T) = I_\ell$), where operator diag$(\cdot)$ chooses the diagonal elements only. Thus, we use the following matrix including feature-feature, feature-label, and label-label

similarities:

$$W = (w_{ij}) = \text{abs}\begin{pmatrix} \frac{1}{n}XX^T & \frac{1}{n}XY^T \\ \frac{1}{n}YX^T & \frac{1}{n}YY^T \end{pmatrix} \in \mathbb{R}^{(d+\ell)\times(d+\ell)},$$

Here, the (i,j)th element is equal to the correlation coefficient ρ_{ij}. We took the absolute value of ρ_{ij} because the sign is irrelevant to the degree of functional relationship.

We minimize

$$\sum_{ab} w_{ab}\|f_a - f_b\|^2 + 2\sum_{am} w_{am}\|f_a - h_m\|^2 + \sum_{mo} w_{mo}\|h_m - h_o\|^2$$

to find low-dimensional representations f_a of feature a and h_m of label m.

(Visualization FL-A)

$$\tilde{f}_a = u_a \in \mathbb{R}^k, \quad \tilde{y}_m = u_{d+m} \in \mathbb{R}^k,$$

where, u_a is ath column of $U_{(k)} \in \mathbb{R}^{k\times(d+\ell)}$ satisfying

$$D^{-1}WU_{(k)}^T = U_{(k)}^T\Lambda_{(k)} \text{ for } W = \text{abs}\begin{pmatrix} \frac{1}{n}XX^T & \frac{1}{n}XY^T \\ \frac{1}{n}YX^T & \frac{1}{n}YY^T \end{pmatrix}.$$

Note that in this visualization the number of samples n affects evenly on features and labels, unlike SL-B.

IV. Proposed Visualization Methods

A. Normalization

We consider two cases of continuous variables and binary variables in either X or Y, or both.

When X (or Y) is of continuous variables, the most natural normalization is centering such that $\mu_x = (1/n)X\mathbf{1} = \mathbf{0}$. The following sphering guarantees the invariance for affine transformations.

When Y (or X) is of binary variables, one of the natural normalization ways is "sum-to-one" normalization such that $Y\mathbf{1} = \mathbf{1}$. In this case, the following sphering does not have a special meaning. In multi-label cases, each output/column y_i can have many one's (labels). Then this normalization guarantees that every label has the same number of samples. Note that the variance and covariance of random variables taking a value in $[0,1]$ is at most one.

B. Proposed methods

In summary, we propose the following two visualization methods:

1) (For simultaneous visualization of samples and labels) Use SL-A with centered and sphered X ($X\mathbf{1} = \mathbf{0}, XX^T = nI_n$) and unprocessed Y. For binary features in X, sum-to-one normalization could be applied instead of centering, with forcibly set $\mu_y = \mathbf{0}$ (without actual centering).
2) (For simultaneous visualization of features and labels) Use FL-A with centered and norm-to-one standardized X ($X\mathbf{1} = \mathbf{0}, \text{diag}(XX^T) = I_d$) and Y ($Y\mathbf{1} = \mathbf{0}, \text{diag}(YY^T) = I_\ell$).

Note that the same information $\frac{1}{n}XY^T$ plays an important role in both methods.

V. Experiments

We dealt with two multi-label datasets taken from Mulan dataset [2]: Scene and Medical. Among them, Scene is a collection of scene images with multi-labels such as "urban", "beach" and "sunset." It has parameters $n = 2407, d = 294, \ell = 6, \ell' = 15$. All the features are continuous. On the other hand, Medical is a collection of clinical texts multi-annotated. All the features are binary/nominal according to if a certain word exists or not. It has parameters $n = 333, d = 1449, \ell = 45, \ell' = 94$.

The result for Scene dataset is shown in Fig. 1. We can see the following: 1) six leading classes can be well classified as a single-label problem in the feature space (Fig. 1(a) with the Volonoi diagram on single labels), 2) the composite classes with two labels (smaller filled circles) are shifted from their component (leading) classes (larger filled circles) (Fig. 1(b)), 3) there is no feature sharing a high correlation with any label (Fig. 1(c)), and 4) some features are redundant (closer-numbered features are indeed closely located) (Fig. 1(d)). The high value of 45.8% in SL-A in the contribution rate [3] indicates the reliability of this low-dimensional visualization.

In multi-label classification problems, it is important to know how composite classes are related to their component classes with single labels. In Fig. 1(b), composite classes moves right from the convex hull of the six component classes. It implies that we need a different classification rule for them, such as to classify them after a constant shift of features. As an analysis of feature-label relationship, we can see in Fig. 1(c)(d) that there are some clusters of features, and some of them are strongly correlated to each other, such as #85 and #92. They can be combined for classification.

The result for Medical dataset is shown in Fig. 2. We can see the following: 1) many labels are difficult to be distinguished (Fig. 2(a) with the Volonoi diagram on single labels), 2) all features are correlated to some labels to some extent (Fig. 2(b)), but not so strong. We also observed that two features associated with words "medical" and "measuring" are located far outside Fig. 2(b),

VI. Kernelized Visualization

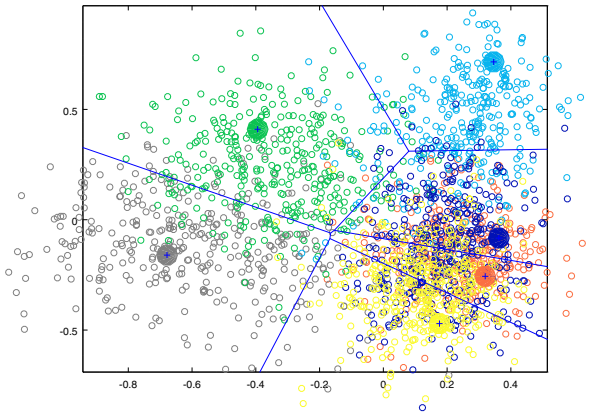
These linear methods are easily extended to nonlinear methods by a kernel trick. When we consider the case in which X is mapped into a higher dimensional sample set $\phi(X)$. Then, the condition by which the linear regression is equivalent to a matrix approximation problem as follows.

Theorem 3. *Let $\Phi(X) \in \mathbb{R}^{q\times n}$ be n samples of X mapped by $\phi(\cdot)$. If $G = \Phi(X)^T\Phi(X) = (K(x_i,x_j)) = nI_n$ [4] and a coefficient matrix A is of linear combinations of $\phi(x_i)$ $(i = 1,2,\ldots,n)$, that is, $A = \Phi(X)B$ for some $B \in \mathbb{R}^{n\times\ell}$, then the linear regression of $Y \in \mathbb{R}^{\ell\times n}$ on $\Phi(X)$ is equivalent to*

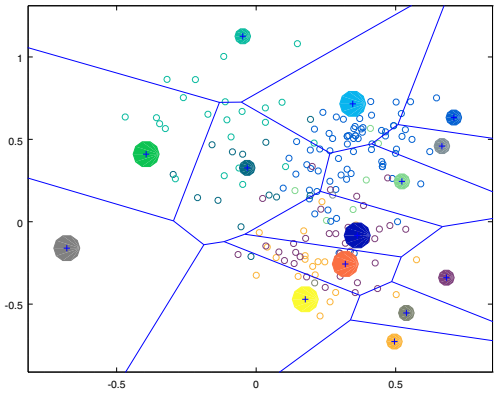
[2] http://mulan.sourceforge.net/datasets-mlc.html

[3] the ratio of the sum of used singular values to the total sum.

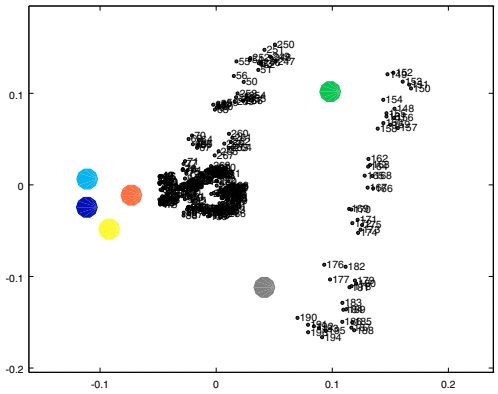
[4] Centering is possible by $G \leftarrow G - \mathbf{n}^{-1}G - G\mathbf{n}^{-1} + \mathbf{n}^{-1}G\mathbf{n}^{-1}$, where $\mathbf{n}^{-1}$ is the $n \times n$ matrix of all $1/n$'s [11]. Sphering is also possible by transformation $\phi(x) \leftarrow \sqrt{n}(\Lambda^g)^{-1}(V^g)^T\Phi(X)^T\phi(x)$, where $G = V^g\Lambda^g(V^g)^T$, $(V^g)^TV^g = V^g(V^g)^T = I_n$, and $\text{rank}(G) = n$ is assumed.



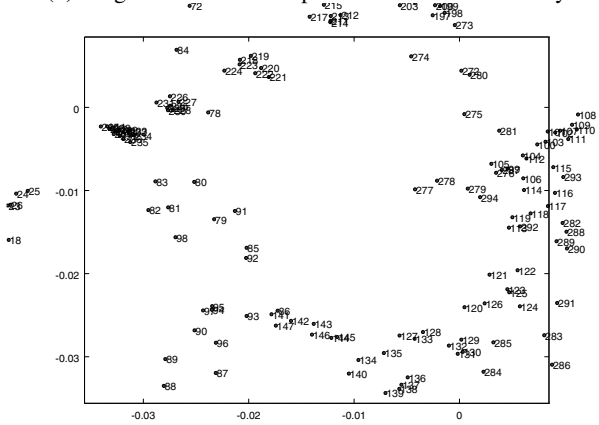
(a) Single-label 6 classes by SL−A (C.R.:45.8%)



(b) Single-label 6 classes plus two-label 8 classes by SL−A



(c) 6 labels and 294 features by FL−A (C.R.:1.5%)



(d) Zoomed some features by FL−A

Fig. 1. Visualization of "Scene" data ($n = 2407, d = 294, \ell = 6, \ell' = 15$) by SL−A and FL−A. C.R. denotes the contribution rate.

an approximation problem in the sense of

$$\arg \min_{A \in \mathbb{R}^{q \times \ell}} \|Y - A^T \Phi(X)\|_F^2 = \arg \min_{B \in \mathbb{R}^{n \times \ell}} \|\frac{1}{n} Y^T - B\|_F^2. \tag{11}$$

From SVD: $Y^T = U \Lambda V^T$, we have the best k-rank approximation

$$B = U_{(k)} \Lambda_{(k)} V_{(k)}^T, \ A = \Phi(X) U_{(k)} \Lambda_{(k)} V_{(k)}^T. \tag{12}$$

Using notation $K(X, x) = (K(x_1, x), \ldots, K(x_n, x))^T$, we have

(Visualization SL−AK)

$$\begin{aligned} \tilde{x} &= \Lambda_{(k)} U_{(k)}^T \Phi(X)^T \phi(x) = \Lambda_{(k)} U_{(k)}^T K(X, x) \in \mathbb{R}^k \\ \tilde{y} &= V_{(k)}^T y \in \mathbb{R}^k, \\ & \quad \text{where } Y^T = U \Lambda V^T. \end{aligned}$$

VII. Discussion

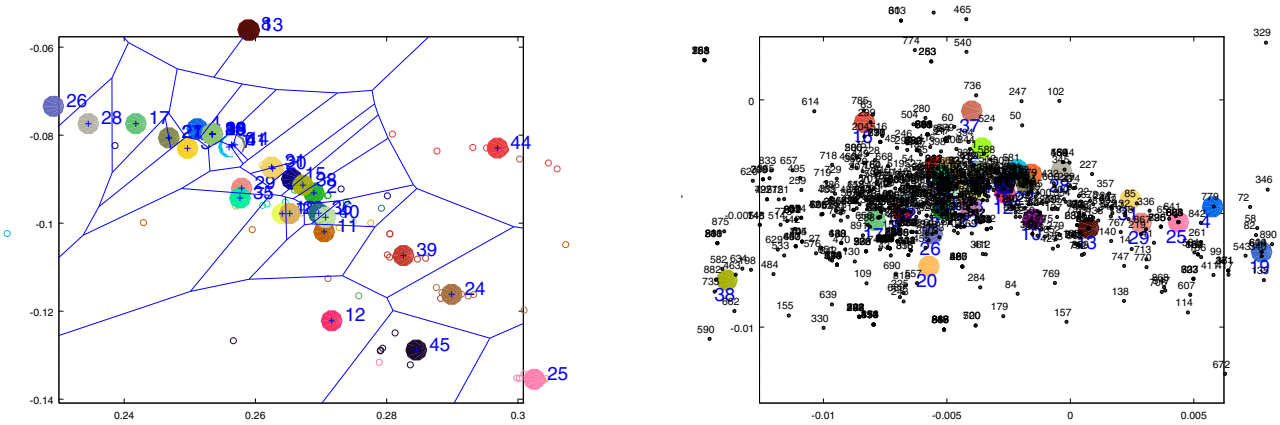
This paper is closely related to Canonical Correlation Analysis (CCA) in which two different kinds of source, X and Y, are considered at the same time, and two low-dimensional mappings $\Psi_x : \mathbb{R}^d \to \mathbb{R}^k$ and $\Psi_y : \{0, 1\}^\ell \to \mathbb{R}^k$ are realized by finding two projections so as to maximize the correlation of them. Indeed, SL−A is identical to a low-rank CCA if X and Y are both sphered, and FL−A is one of embeddings using the two mappings of CCA. Nevertheless, many studies have shown either Ψ_x or Ψ_y only. This is because it is meaningless to use the two mappings at the same time without considering under which condition the regression and CCA are identified. Sun *et al.* [12] showed already an important theorem on it, but ours is more understandable because we have given a necessary and sufficient condition, $XX^T = nI_d$. Some papers carried out a simultaneous mapping under stronger conditions such as both X and Y are centered and sphered, but they have not paid a special attention to such conditions.

Returning to Theorem 1, let us examine the significance of amount XY^T/n. It is related to many standard technologies under the sphered condition on X. First, it is the coefficient matrix of a linear regression of Y on X [5] because

$$\hat{y}^T = x^T A = x^T \left((XX^T)^{-1} XY^T \right) = x^T \left(\frac{1}{n} XY^T \right).$$

Next, suppose that both X and Y are centered to $\mu_x = \mu_y = \mathbf{0}$. Then this amount is the expected mean of y given x,

[5] This is the proof that $XX^T = nI_d$ is a necessary condition.



(a) Single-label some classes by SL−A (C.R.:16.1%)

(b) Single-label some classes and features by FL−A

Fig. 2. Visualization of "Medical" data ($n = 333, d = 1449, \ell = 45, \ell' = 94$) by SL−A and FL−A. C.R. denotes the contribution rate.

assuming $(x^T, y^T)^T$ obeys a Gaussian. This can be shown as

$$\begin{aligned} \mathbb{E}\, y|x &= \mu_y + \Sigma_{yx}\Sigma_{xx}^{-1}(x - \mu_x) \\ &= (\frac{1}{n}YX^T)(\frac{1}{n}XX^T)^{-1}x = \frac{1}{n}YX^Tx. \end{aligned}$$

In the viewpoint of partial least square techniques, this amount appears in the orthonormal partial least square [5] and its two-stage implementation [6] because

$$\begin{aligned} &\min_{U,V} \|(XX^T)^{-1/2}XY^T - U\Lambda_{(k)}V^T\| \\ = &\min_{U,V} \|\frac{1}{\sqrt{n}}XY^T - U\Lambda_{(k)}V^T\|. \end{aligned}$$

In addition, it can be shown that above U gives the solution of Fisher's LDA in single-label cases.

Another viewpoint on this amount is possible. We have already pointed out that

$$\|Y - A^TX\|_F^2 = \|Y - (1/n)YX^TX\|_F^2$$

without low-rank restriction. Here term X^TX can be kernelized to $G = (K(x_i, x_j))$. Then our requirement $G = nI_n$ in Theorem 3 derives

$$\|Y - A^T\Phi(X)\|_F^2 = \|Y - (1/n)YG\|_F^2 = 0.$$

That is, term X^TX and its kernelized version $(K(x_i, x_j))$ show how individual samples are mutually independent of others, and in the extreme case of $(K(x_i, x_j)) = (n\delta_{ij})$, the perfect reconstruction is realized. In that case, every training sample x_i is mapped on one of $\tilde{y}_m (m = 1, 2, \ldots, \ell')$ where ℓ' is the number of distinct label subsets in Y.

VIII. Conclusion

We have shown three algorithms for simultaneous visualization of samples and labels and one algorithm for simultaneous visualization of features and labels. The proposed visualization methods allow fast visual assessment of feature's information content and their discriminability in addition to assessment of multi-labels. In addition, we showed that XY^T/n plays a central role in these algorithms. Unfortunately their practical usefulness has confirmed only partially. We will deepen the potential and explore their applications.

Acknowledgment

This work was partially supported by JSPS KAKENHI Grant Number 15H02719.

References

[1] T. Hastie, R. Tibshirani and J. Friedman, "The Elements of Statistical Learning" (2nd ed.), Springer New York Inc., 2009.

[2] S.T. Roweis and L.K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding." Science, 290-5500(2000), 2323–2326.

[3] Geng, X., Zhan, D. C., and Zhou, Z. H. "Supervised nonlinear dimensionality reduction for visualization and classification." IEEE Transactions onSystems, Man, and Cybernetics, Part B: Cybernetics, 35-6(2005), 1098-1107.

[4] Kimura et al., "Simultaneous Nonlinear Label-Instance Embedding for Multi-label Classification." submitted to S+SSPR2016.

[5] K. Worsley *et al.* A. Evans, "Characterizing the response of PET and fMRI data using multivariate linear models." Neuroimage, 6(4):305319, 1997.

[6] L. Sun, B. Ceran, and J. Ye, "A scalable two-stage approach for a class of dimensionality reduction techniques." In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 313322, 2010.

[7] E. Kokiopoulou and J. Chen and Y. Saad, "Trace optimization and eigenproblems in dimension reduction methods." Numerical Linear Algebra with Applications, 18-3 (2011), 565-602.

[8] Adi Ben-Israel and Thomas N.E. Greville, Generalized Inverses: Theory and Applications (2nd ed.), Springer-Verlarg, New York, 2003.

[9] M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation." Neural computation, 15(6):13731396, 2003.

[10] A. Ben-Israel and T. N.E. Greville, "Generalized Inverses: Theory and Applications." (2nd ed.), Springer-Verlarg, New York, 2003.

[11] B. Schölkopf and A. J. Smola, "Learning with Kernels." (p.431), MIT press, 2002.

[12] L. Sun *el.*, "Cannonical COrrelation Analysis for Multilabel Classification: A Least-Square Formulation, Extensions, and Analysus, IEEE Trans. on PAMI, 33-1(2011), 194–200.