Supplementary Information

A Self-Training Subspace Clustering Algorithm under Low-Rank Representation for Cancer Classification on Gene Expression Data

Chun-Qiu Xia[1], Ke Han[1], Yong Qi[1], Yang Zhang[2,*] and Dong-Jun Yu[1,2,*]

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Xiaolingwei, Nanjing 210094, China

[2] Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, Michigan 48109, USA

[*] Address correspondence to Y. Zhang at zhng@umich.edu or D. J. Yu at njyudj@njust.edu.cn

Tel: +86-025-84316190

Fax: +86-025-84315960

1. The detailed composition of the benchmark datasets

Table S1. The detailed composition of the R-GCM and MBD datasets.

Dataset	No. of Genes	Cancer Type / Subgroup	No. of Samples	Cancer Type / Subgroup	No. of Samples
R-GCM	11,370	Breast adenocarcinoma (BR)	12	Uterus adenocarcinoma (UT)	10
		Prostate adenocarcinoma (PR)	14	Leukemia (LE)	30
		Lung adenocarcinoma (LU)	12	Renal cell carcinoma (RE)	11
		Colorectal adenocarcinoma (CO)	12	Pancreas adenocarcinoma (PA)	11
		Lymphoma (LY)	22	Ovarian adenocarcinoma (OV)	12
		Bladder transitional cell carcinoma (BL)	11	Pleural mesothelioma (MS)	11
		Melanoma (ML)	10	Central nervous system (CNS)	20
MBD	54,675	WNT-subgroup (WNT)	8	subgroup-3 (G3)	16
		SHH-subgroup (SHH)	10	subgroup-4 (G4)	39

2. Low-Rank representation of high-dimension data

Low-rank representation (LRR) was first proposed by Liu et al. [1], which is a matrix decomposition algorithm to segment high-dimensional data into subspaces for reducing its dimensions. Let $\mathbf{x} = [x_1, x_2, \cdots, x_i, \cdots, x_d]^{\mathrm{T}} \in \mathbf{R}^d$ be a *d*-dimensional feature vector of a sample (e.g., gene expression data in this study), where x_i is the *i*-th feature component. A set of *n* samples can then be formulated by a data matrix $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n] \in \mathbf{R}^{d \times n}$, each column of which is a sample. $\mathbf{X}$ can

be decomposed as:

$$\mathbf{X}=\mathbf{AZ} \tag{S1}$$

where $\mathbf{A}=[\mathbf{a}_1,\mathbf{a}_2,\cdots,\mathbf{a}_m]\in\mathbf{R}^{d\times m}$ is a dictionary matrix which consists of m bases, and $\mathbf{Z}=[\mathbf{z}_1,\mathbf{z}_2,...,\mathbf{z}_n]\in\mathbf{R}^{m\times n}$ is a coefficient matrix. Based on this decomposition, each sample $\mathbf{x}_i$ can be represented by the linear combination of the m bases in dictionary $\mathbf{A}$ with the corresponding coefficient $\mathbf{z}_i$. Hence, $\mathbf{z}_i$ is termed as the representation of $\mathbf{x}_i$.

To capture the global structure of the data $\mathbf{X}$, low-rankness is used as the criterion to guide the decomposition of $\mathbf{X}$ by solving the problem of Eq. (S2) [1]:

$$\begin{cases}\min\limits_{\mathbf{Z}} rank(\mathbf{Z})\\ s.t.,\ \mathbf{X}=\mathbf{AZ}\end{cases} \tag{S2}$$

The optimal solution $\mathbf{Z}^*$ of Eq. (S2) is termed as the "lowest-rank representations" of data $\mathbf{X}$ with respect to dictionary $\mathbf{A}$.

In real-world application scenarios, the data are often noisy, redundant and have outliers. Extracting an additional sparse matrix $\mathbf{E}=[\mathbf{e}_1,\mathbf{e}_2,\cdots,\mathbf{e}_n]\in\mathbf{R}^{d\times n}$ from $\mathbf{X}$ would help to reduce noises and recover the intrinsic global structure of the data [1]. We assume that matrix $\mathbf{E}$ should be sparse because noise often only affects a part of data. Hence, the problem of Eq. (S2) can be reformulated as:

$$\begin{cases}\min\limits_{\mathbf{Z},\mathbf{E}} rank(\mathbf{Z})+\lambda\left\|\mathbf{E}\right\|_0\\ s.t.,\ \mathbf{X}=\mathbf{AZ}+\mathbf{E}\end{cases} \tag{S3}$$

where $\left\|\mathbf{E}\right\|_0$ means the l_0-norm of matrix **E**, which guarantees the sparsity of the matrix, and λ is a parameter to balance the effects of the two parts. Problem in Eq. (S3) is an optimization problem, which is difficult to solve due to its non-convexity. Here, we replace the minimizations of the rank of $\mathbf{Z}$ and the l_0-norm of $\mathbf{E}$ with that of the nuclear norm of $\mathbf{Z}$ and the $l_{2,1}$-norm of $\mathbf{E}$, respectively, following Ref. [1]. Hence, the optimization problem of Eq. (S3) can be reformulated as

$$\begin{cases}\min\limits_{\mathbf{Z},\mathbf{E}} \left\|\mathbf{Z}\right\|_*+\lambda\left\|\mathbf{E}\right\|_{2,1}\\ s.t.,\ \mathbf{X}=\mathbf{AZ}+\mathbf{E}\end{cases} \tag{S4}$$

where $\left\|\mathbf{Z}\right\|_*=\sum_i\sigma_i(\mathbf{Z})$ denotes the nuclear norm of $\mathbf{Z}$, $\sigma_i(\mathbf{Z})$ means the i-th singular value of $\mathbf{Z}$, and $\left\|\mathbf{E}\right\|_{2,1}=\sum_{j=1}^{n}\sqrt{\sum_{i=1}^{d}\left([\mathbf{E}]_{ij}\right)^2}$ denotes the $l_{2,1}$-norm of $\mathbf{E}$.

According to self-expressive scheme, the data matrix itself can serve as the dictionary, which not only contributes to subspace segmentation, but also simplifies the calculation [2]. Accordingly, the

Eq. (S4) can be rewritten in the form of

$$\begin{cases} \min\limits_{\mathbf{Z},\mathbf{E}} \|\mathbf{Z}\|_* + \lambda \|\mathbf{E}\|_{2,1} \\ s.t.,\ \mathbf{X} = \mathbf{XZ} + \mathbf{E} \end{cases} \tag{S5}$$

The optimal solution $\mathbf{Z}^*$ of Eq. (S5) is the final lowest-rank representation of the original data matrix $\mathbf{X}$, which will be used for the subsequent computation in this study.

3. Configuration parameters

(1) λ ***for LRR.*** The λ in LRR is a regularizing parameter which balances the effects of $\|\mathbf{Z}\|_*$ and $\|\mathbf{E}\|_{2,1}$ (cf. Eq. (10)). We applied a grid search strategy to optimize the value of λ, and found that the proposed method can achieve the robust satisfactory performance when $\lambda \in \left[2^{-3}, 2^4\right]$.

(2) Distance metrics for clustering. The selection of distance metric is crucial for the success of a clustering algorithm, and a suitable distance metric will do great help to achieve good clustering performance. Jaskowiak et al. [3] presented several guidelines, which are intrinsically empirical-based, on how to choose distance metric for gene expression data. Hence, we also exploited several distance metrics, including Euclidian, Minkowshi, Cosine, Pearson, and Rank-Magnitude [4], on the benchmark datasets. According to our local testing results, it was found that the Minkowshi metric as defined in Eq. (S11) is a better choice.

$$\text{Minkowshi}(\mathbf{a},\mathbf{b}) = \left(\sum_{i=1}^{n} |a_i - b_i|^{\xi} \right)^{1/\xi} \tag{S6}$$

Nevertheless, there is still a parameter ξ in Minkowshi distance metric. We optimized the value of ξ by performing a grid search strategy on each of the two benchmark datasets and found that $\{\xi_{\mathbf{Z}} = 2^3, \xi_{\mathbf{E}} = 2^0\}$ and $\{\xi_{\mathbf{Z}} = 2^2, \xi_{\mathbf{E}} = 2^2\}$ are best choices for R-GCM and MBD, respectively.

(3) Parameters for SVM, Label Propagation, and Semi-PNMF. SVM [5] is a typical supervised classification/regression algorithm, which has been widely used in bioinformatics fields including cancer classification because of its superior performance. Radial basis function (RBF) is chosen as the kernel function. The other two parameters, i.e., the regularization parameter γ and the kernel width parameter σ of RBF, are set according to the optimization results from a grid search strategy with a 10-fold cross validation. Label propagation [6] is a commonly-used semi-supervised learning model, which is implemented using a RBF kernel function in this study. The kernel width parameter σ of RBF and the clamping factor α in Label Propagation, are set using the strategy similar to that in SVM. Semi-PNMF [7] is a recently proposed semi-supervised learning model that is

specifically designed for cancer classification on gene expression data. The two parameters in Semi-PNMF, i.e., regularization constant α and trade-off factor β, are optimized by using the method suggested in reference [7]. The optimized parameter pairs of (α,β) Semi-PNMF are $(\alpha=2,\beta=0.0001)$ and $(\alpha=0.2,\beta=0.01)$ on R-GCM and MBD datasets, respectively.

4. Traditional self-training method

In machine learning field, self-training refers to a type of semi-supervised method which can utilize both the labeled and unlabeled data [8]. Traditional self-learning works as follows: For a given dataset with labeled and unlabeled data, it first trains a learning model (e.g., classification or clustering model) based on the labeled data; Then, the trained model is used to predict the labels of the unlabeled data, and those unlabeled data, whose labels are predicted with high confidence (score), will be selected and added to the training subset; This practice continues until the entire dataset is labeled. Self-training method is particularly suitable for processing biological data, where there often exists large volume of unlabeled data due to the high cost of annotation [9]. However, traditional self-training method has the potential to reinforce model mistakes, termed as mistake-reinforcement, if falsely predicted data are selected and added into the training subset during the data selection procedure in each iteration [10-12].

REFERENCES

[1] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 663-670.

[2] E. Elhamifar and R. Vidal, "Sparse Subspace Clustering," *CVPR: 2009 IEEE Conference on Computer Vision and Pattern Recognition,* vol. 1-4, pp. 2782-2789, 2009.

[3] P. A. Jaskowiak, R. J. G. B. Campello, and I. G. Costa, "On the selection of appropriate distances for gene expression data clustering," *BMC Bioinformatics,* vol. 15, p. S2, Jan 24 2014.

[4] R. J. G. B. Campello and E. R. Hruschka, "On comparing two sequences of numbers and its applications to clustering analysis," *Information Sciences,* vol. 179, pp. 1025-1039, 2009.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning,* vol. 20, pp. 273-297, 1995.

[6] X. Zhu and Z. Ghahramani, "Learning from Labeled and Unlabeled Data with Label Propagation," Carnegie Mellon University: School of Computer Science, Pittsburgh, PA, USA CMU-CALD-02-107, Jun 2002.

[7] X. Zhang, N. Guan, Z. Jia, X. Qiu, and Z. Luo, "Semi-Supervised Projective Non-Negative Matrix Factorization for Cancer Classification," *PLoS One,* vol. 10, p. e0138814, Sep 22 2015.

[8] O. Chapelle, B. Scholkopf, and A. Zien, "Semi-Supervised Learning (Chapelle, O. et al., Eds.; 2006)[Book reviews]," *IEEE Transactions on Neural Networks,* vol. 20, pp. 542-542, Feb 2009.

[9] R. Ibrahim, N. A. Yousri, M. A. Ismail, and N. M. El-Makky, "miRNA and Gene Expression based Cancer Classification using Self-Learning and Co-Training Approaches," *2013 IEEE International Conference on*

Bioinformatics and Biomedicine (BIBM), pp. 495-498, 2013.

[10] X. R. Zhao, N. Evans, and J. L. Dugelay, "Semi-Supervised Face Recognition with Lda Self-Training," *2011 18th IEEE International Conference on Image Processing,* pp. 3041-3044, 2011.

[11] K. Prokopiou, E. Kavallieratou, and E. Stamatatos, "An Image Processing Self-Training System for Ruling Line Removal Algorithms," *2013 18th International Conference on Digital Signal Processing (DSP),* pp. 1-6, 2013.

[12] X. Zhu, "Semi-Supervised Learning Literature Survey," *Computer Science,* vol. 37, pp. 63-77, 2008.