Diagnosis-Guided Multi-modal Feature Selection for Prognosis Prediction of Lung Squamous Cell Carcinoma

Wei Shao[1], Tongxin Wang[2], Zhi Huang[5], Jun Cheng[3], Zhi Han[1], Daoqiang Zhang[4(✉)], and Kun Huang[1(✉)]

[1] School of Medicine, Indiana University, Indianapolis, IN 46202, USA
kunhuang@iu.edu
[2] Department of Computer Science, Indiana University Bloomington, Bloomington, IN 47405, USA
[3] School of Biomedical Engineering, Shenzhen University, Shenzhen 518073, China
[4] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
dqzhang@nuaa.edu.cn
[5] School of Electrical and Computer Engineering, Purdue University, West Lafayette, IN 47907, USA

Abstract. The existing studies have demonstrated that the integrative analysis of histopathological images and genomic data can hold great promise for survival analysis of cancers. However, direct combination of multi-modal data may bring irrelevant or redundant features that will harm the prognosis performance. Therefore, it has become a challenge to select informative features from the derived heterogeneous data for survival analysis. Most existing feature selection methods only utilized the collected multi-modal data and survival information to identify a subset of relevant features, which neglect to use the diagnosis information to guide the feature selection process. In fact, the diagnosis information (*e.g.,* TNM stage) indicates the extent of the disease severity that are highly correlated with the patients' survival. Accordingly, we propose a diagnosis-guided multi-modal feature selection method (DGM2FS) for prognosis prediction. Specifically, we make use of the task relationship learning framework to automatically discover the relations between the diagnosis and prognosis tasks, through which we can identify important survival-associated image and eigengenes features with the help of diagnosis information. In addition, we also consider the association between

This work was supported in part by the Indiana University Precision Health Initiative to ZH and KH, the National Natural Science Foundation of China (61876082, 61861130366, 61703301) to DZ, and Shenzhen Peacock Plan (KQTD20160531120514 97) to JC.

Electronic supplementary material The online version of this chapter (https://doi.org/10.1007/978-3-030-32251-9_13) contains supplementary material, which is available to authorized users.

D. Shen et al. (Eds.): MICCAI 2019, LNCS 11767, pp. 113–121, 2019.
https://doi.org/10.1007/978-3-030-32251-9_13

the multi-modal data and use a regularization term to capture the correlation between the image and eigengene data. Experimental results on a lung squamous cell carcinoma dataset imply that incorporating diagnosis information can help identify meaningful survival-associated features, by which we can achieve better prognosis prediction performance than the conventional methods.

1 Introduction

One of the long-term goals of cancer research is to identify prognostic factors that affect patients' survival time, which in turn allows clinicians to make early decision on treatment [1]. So far, many biomarkers have been shown to be sensitive to the prognosis of cancers [2]. For example, quite a number of cancer prognosis models are based on the histopathological images, since they reveal the morphological attributes of cells that are closely associated with the progress of cancer disease. Moreover, it is known that mutations in genes can cause cancer by accelerating cell division rates. Accordingly, many researchers also use the genomic data such as gene expression signatures to drive cancer prognosis.

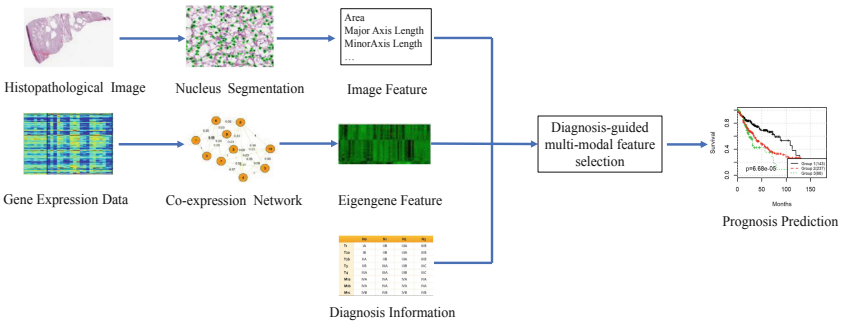

Fig. 1. The flowchart of the proposed method.

In many medical applications, it is common to acquire multiple biomarkers to predict the disease status more accurately. Recently, some researchers also integrated the histopathological imaging and gene-level data to predict the clinical outcome of cancer patients [3,4]. For instance, Yuan *et al.* [3] combined the pathological image data with copy number variation data to devise an integrated predictor of survival for breast cancer patients. Yao *et al.* [4] developed a novel correlational survival prediction framework for the integration of pathological images and genomic data on different cancer cohorts. All these results indicate that using multiple biomarkers can reveal hidden information that are overlooked by only using imaging or genomic data, and thus can better predict patients' clinical outcome.

Although integrating imaging and genomic features can achieve promising results, simply combining them may bring the problem of curse of dimensionalities. Thus, feature selection, which can be considered as the identification of key biomarkers, is commonly used to select useful features. Specifically, the studies in [2,3] firstly put the features from different modalities together and then applied the traditional single-modality feature selection method (*i.e.,* LASSO) to discover the key components that affect the patients' survival. Shao *et al.* [5] exploited and utilized the correlation across multi-modal data to select important imaging and genomic biomarkers related to cancer prognosis. However, these feature selection algorithms are only based on the collected multi-modal data and the survival information, which neglect to take the prior diagnosis information into consideration. As a matter of fact, the diagnosis information (*i.e.,* TNM stage) could also provide guidance to help predict the clinical outcome of patients. For example, the patients who are in stage III suffer from more aggressive cancers than those in stage I, and thus generally have higher survival risks.

Based on the above considerations, in this paper, we propose a diagnosis-guided multi-modal feature selection method (DGM2FS) to discover relevant imaging and genomic risk factors for survival analysis. Specifically, based on the task relationship learning approach, our method can automatically derive the correlation between the diagnosis and prognosis tasks to help identify survival-associated features. In addition, we also consider the association between multi-modal data by adding a regularization term that is capable of capturing the inter-correlation between the selected imaging and genomic components. The experimental results on a public available lung squamous cell carcinoma (LUSC) dataset demonstrate that the proposed method outperforms conventional methods in terms of prognosis prediction.

Table 1. Demographics and clinical characteristics of the LUSC dataset.

Characteristics	Summary	Characteristics	Summary
Patients:		Stage:	
Censored	258	Stage I	218
Non-censored	188	Stage II	151
Age (Y):	58.3 ± 13.1	Stage III	77
Follow-up (M):	47.4 ± 23.2		

2 Method

We summarize our framework in Fig. 1, which consists of the following 3 steps, *i.e.,* feature extraction, the proposed diagnosis-guided multi-modal feature selection method (DGM2FS), and prognostic prediction. We will firstly provide the description of the dataset used in this study.

Dataset: The Cancer Genome Atlas (TCGA) project has generated genomic and imaging data for thousands of tumour samples across more than 30 types of cancers. In this study, we test our method on the lung squamous cell carcinoma (LUSC) cohort derived from TCGA. Specifically, the LUSC dataset contains the H&E stained whole-slide images and gene expression data for 446 patients. For each patient, the corresponding survival information (*i.e.,* survival status, survival time) and the TNM stage information are all available. We show the details of the cohort information in Table 1, where censored patients mean that these patients did not suffer the outcome of death events during the follow-up period, while the non-censored category refers to the patients whose survival information are accurate from diagnosis to death.

Feature Extraction: For whole-slide images of each patient, 2−8 regions of interest (ROI) of size 3K by 3K are extracted at first. Then, we apply an unsupervised method introduced in [6] to segment the nuclei from these extracted patches. Next, for each segmented nucleus, we extract seven cell-level features, including nuclei area (denoted as area), the major and minor axis length of cell nucleus, the ratio of major axis length to minor axis length (major, minor, and ratio), and mean, maximum, and minimum distances (distMean, distMax, and distMin) to its neighbouring nuclei. Finally, for each cell-level feature, a 10-bin histogram and five statistics (*i.e.,* mean, SD, skewness, kurtosis, and entropy) are used to aggregate the cell-level features into patient-level features, and thus a 105-dimensional imaging feature can be obtained for each patient. As to gene expression data, to overcome the large number of genes which poses a challenge to the statistical analysis, we use the co-expression network analysis algorithms [7] to cluster genes into co-expressed modules and summarize each module into an eigengene using singular value decomposition, and this process yields to 58-dimensional eigengene features.

Multi-modal Feature Selection for Prognosis Prediction: For the derived imaging and eigengene features, we firstly introduce the multi-modal feature selection (M2FS) algorithm for identifying survival-associated features without the guidance of diagnosis information. Specifically, let $X = [x_1, x_2, ..., x_N]^T = [X^H, X^G] \in R^{N\times(p+q)}$. Here, $X^H \in R^{N\times p}$ and $X^G \in R^{N\times q}$ correspond to the histopathological imaging data and eigengene data. N is the number of the patients, and p and q are the feature number of imaging data and eigengene data, respectively. We use a triplet (x_i, t_i, δ_i) to represent each observation in survival analysis, where $x_i \in R^{p+q}$ is the patient vector, t_i is the observed survival time, and δ_i is the censoring indicator. Here, $\delta_i = 1$ or $\delta_i = 0$ indicates a non-censored or a censored instance, respectively. Then, the objective function of M2FS is:

$$\begin{aligned} \min_{w_s} \sum_{i=1}^{N} l_S(x_i) + \alpha \left\| X^H w_S^H - X^G w_S^G \right\|_2^2 + \beta_S^H \left\| w_S^H \right\|_1 + \beta_S^G \left\| w_S^G \right\|_1 \\ s.t. \quad \left\| X^H w_S^H \right\|_2^2 \leq 1; \left\| X^G w_S^G \right\|_2^2 \leq 1; \end{aligned} \tag{1}$$

In the M2FS model, $l_S(x_i) = c_i \left\| x_i w_S - t_i \right\|_2^2$ is a weighted regression term, where $w_s = [w_S^H, w_S^G] \in R^{p+q}$ and c_i is the weight of the survival regression loss for each patient defined as follows:

$$c_i = \begin{cases} 1 & if \ \ \delta = 1 \\ 0 & if\ \delta = 0 \ \ and \ \ x_i w_S - t_i > 0 \\ \sigma & if\ \delta = 0 \ \ and \ \ x_i w_S - t_i \leq 0 \end{cases} \tag{2}$$

where $c_i = 1$ if x_i is a non-censored patient. By considering that the actual survival time for a censored instance should be larger than its observed time, we define c_i as 0 if $x_i w_S - t_i \geq 0$. Otherwise, we let c_i equal to a constant σ that is larger than 1 since the difference between actual survival time and the estimated survival time is indeed greater than $t_i - x_i w_S$. The second term in Eq. (1) is used to minimize the distance between the projections of imaging *i.e.,* $X^H w_S^H$ and genomic *i.e.,* $X^G w_S^G$ data, so that their inter-correlations can be captured. The third and fourth L1-norm terms are used to select a small number of image and eigengene features for the following prognosis prediction.

Diagnosis-Guided Multi-modal Feature Selection: In the above M2FS method, we only utilize the multi-modal data and the survival information for prognosis prediction, which overlooks the available diagnosis information (*i.e.,* TNM stage) is also correlated with patients' survival. To address this problem, we propose to incorporate the diagnosis information and use the task relationship learning term [8] to automatically discover the relationship between the diagnosis and prognosis tasks to boost the prognosis prediction performance. The task relationship learning term is defined as $tr(WM^{-1}W^T)$ with $M \geq 0$ and $tr(M) = 1$. Here, $W = [w_D, w_S] \in R^{(p+q)\times 2}$, where $w_D = [w_D^H, w_D^G]^T$ and $w_S = [w_S^H, w_S^G]^T$ correspond to the linear discriminant functions for the diagnosis and prognosis tasks on the multi-modal data, respectively. M^{-1} denotes the inverse of the matrix $M \in R^{2\times 2}$, and M is defined as a task covariance matrix that will benefit learning on W by automatically inducing the correct relationship M between different tasks. The constraint term $M \geq 0$ is used to restrict M as positive semidefinite matrix, and $tr(M) = 1$ is used to penalize the complexity of M. By incorporating the above relationship induced term, the objective function of the proposed diagnosis-guided multi-modal feature selection (*i.e.,* DGM2FS) method can be formulated as:

$$\begin{aligned} \min_{W,M} & \sum_{k\in\{S,D\}} \sum_{i=1}^{N} l_k(x_i) + \alpha \sum_{k\in\{S,D\}} \left\| X^H w_k^H - X^G w_k^G \right\|_2^2 \\ & + \sum_{k\in\{S,D\}} \sum_{j\in\{H,G\}} \beta_k^j \left\| w_k^j \right\|_1 + \gamma tr(WM^{-1}W^T) \\ s.t. \quad & M \geq 0, tr(M) = 1 \end{aligned} \tag{3}$$

In comparison with the above M2FS model (shown in Eq. (1)), the first term of Eq. (3) further incorporates the empirical loss for the diagnosis task *i.e.,*

$l_D(x_i) = \|x_i w_D - y_i\|_2^2$, where y_i indicates the categorical TNM stage for patient i. The second term is used to capture the correlation between the imaging and eigengene data for both diagnosis and prognosis tasks. The third term enforces some elements of w_k^j to be zero, and thus can select important features for different tasks. In what follows, we will develop an efficient optimization algorithm to solve the objective function defined in Eq. (3).

Optimization: We adopt an alternating strategy to optimize W and M in the proposed DGM2FS model. Specifically, given a fixed M, we define $M^{-1} = [M_{DD}, M_{DS}; M_{DS}, M_{SS}] \in R^{2\times 2}$. Then, the optimization problem for W is:

$$\begin{aligned}\min_W \left\|C(X^H w_S^H + X^G w_S^G - T)\right\|_2^2 + \left\|X^H w_D^H + X^G w_D^G - Y\right\|_2^2 \\ + \alpha \sum_{k\in\{S,D\}} \left\|X^H w_k^H - X^G w_k^G\right\|_2^2 + 2rM_{DS} \sum_{j\in\{H,G\}} (w_D^j)^T w_S^j \\ + \sum_{k\in\{S,D\}} \sum_{j\in\{H,G\}} (\beta_k^j \left\|w_k^j\right\|_1 + rM_{kk}(w_k^j)^T w_k^j)\end{aligned} \tag{4}$$

Where, $T = [t_1, t_2, ..., t_N]^T \in R^N$ and $Y = [y_1, y_2, ..., y_N]^T \in R^N$ indicate the recorded survival time and TNM stage information for the patients in the training set. $C \in R^{N\times N}$ is a diagonal matrix, with the k-th element as c_i defined in Eq. (2). It is worth noting that, Eq. (4) is convex with respect to each $w_k^j (k \in \{S, D\}, j \in \{H, G\})$ in W, and thus can be solved alternatively to obtain the optimal W. After W is determined, we follow the method in [8] to get the closed-formed solution for $M = (W^T W)^{\frac{1}{2}}/tr((W^T W)^{\frac{1}{2}})$. Finally, we renew the weight of each instance *i.e.,* $c(i)$ according to Eq. (2). The above steps will repeat until W, M converge to fixed values.

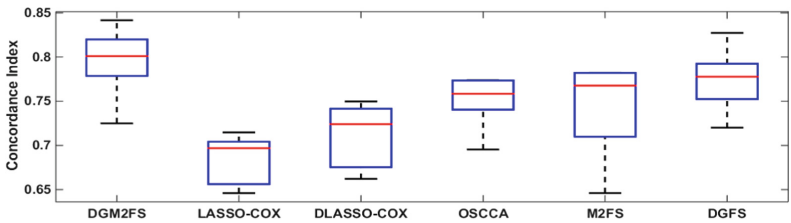

Fig. 2. Comparison of different feature selection methods on Concordance Index.

Prognostic Prediction: We build the Cox model [4] for prognosis prediction. Specifically, we firstly divide all patients into 5 folds, with 4 folds used for training and the remaining for testing, then the Cox proportional hazards model is built on the selected features in the training set, through which we calculate the Concordance Index (CI) [4] that measures the fraction of all pairs of patients

whose survival risks are correctly ordered on the testing set to evaluate the performance of prognosis prediction. The CI value ranges from 0 to 1, where larger CI value means the better prediction performance and vice versa.

3 Experimental Results

Experimental Settings: The parameters $\alpha, \beta_k^j (k \in \{S, D\}, j \in \{H, G\})$ and r in the proposed DGM2FS are tuned from $\{0.5, 1, 1.5\}$,$\{0.1, 0.5, 1\}$ and $\{1, 1.5, 2\}$, respectively. The parameter σ in Eq. (2) is fixed as 1.5.

Results and Discussion: We compare DGM2FS with the following baseline methods by the measurement of CI. (1) LASSO-Cox [2]: use the LASSO method for variable selection in the Cox model. (2) DLasso-Cox: add the diagnosis information as a feature to the multi-modal data, then use the LASSO-Cox model for feature selection. (3) OSCCA [5]: a multi-modal feature selection method for survival analysis without the guidance of diagnosis knowledge. (4) M2FS: a variant of DGM2FS (shown in Eq. (1)), which neglects to take the diagnosis information into consideration. (5) DGFS: a variant of DGM2FS, which miss the second term in Eq. (3) that can capture the correlation between imaging and genomic data. The results are shown in Fig. 2. As can be seen from Fig. 2, DGM2FS and its variant DGFS achieve higher CI value than the competing methods. These results clearly demonstrate that the incorporation of the diagnosis information (*i.e.,* TNM stage) under task relationship learning framework can help improve the prognosis performance. In addition, we observe that the DGM2FS model could provide better prognostic prediction (0.795 ± 0.043) than the DGFS algorithm (0.769 ± 0.039), which also shows the advantage of taking the correlation among different modalities into account for feature selection.

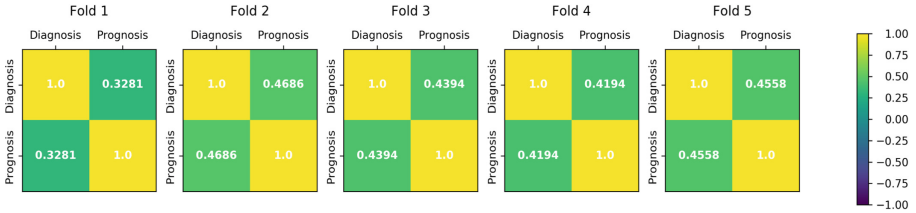

Fig. 3. The correlation coefficient matrix between the diagnosis and prognosis tasks for each fold of cross-validation learned by DGM2FS.

In addition, we visualize the learned correlation coefficient matrix (*i.e.,* calculated from M in Eq. (3)) between the diagnosis and prognosis tasks in Fig. 3. From Fig. 3, we observe that the diagnosis and prognosis tasks are *positively correlated* in each fold of cross-validation, which again verify that adding the diagnosis information could improve the performance of prognosis prediction.

Since it is of great importance to identify the biomarkers that affect prognosis prediction. We focus on the biological significance of the selected imaging and eigenegene features that are appeared at least four times in the 5-fold cross-validation. Specifically, our method selects 6 image features including *major_bin6, distMean_bin2, distMin_kurtosis, distMin_bin3, distMean_entropy* and *distMin_entropy*, where most of them are related to the distance among segmented cells. The cells with smaller neighboring distance (*i.e., disMean_bin2 and distMin_bin3*) usually correspond to the cancer cells or lymphocytes that cluster together in the tissue images, which is generally considered to be the key factors affecting the survival of lung cancer patients [9]. As to genomic features, three eigengenes *i.e., eigengene 10, eigengene 25, eigengene 29* are identified (details are shown in Supplementary). The enrichment analysis on eigengene 10 shows that its corresponding module contains 30 genes, and 14 of them are enriched with the biological process of immune response, which is consistent with the existing study [10] that the immune response plays an important role in the development of lung cancer.

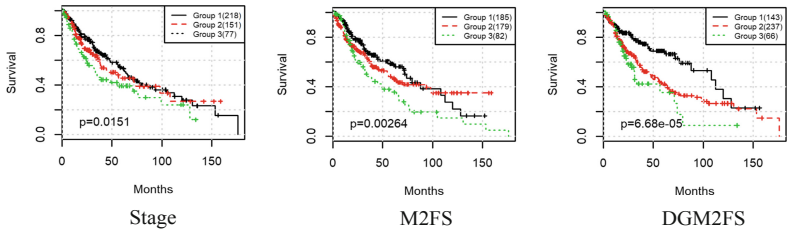

Fig. 4. Comparison of different feature selection methods for patient stratification.

Finally, we also compare the prognostic power of different approaches by stratifying cancer patients into 3 subgroups with different predicted outcomes, with the experimental results shown in Fig. 4. Specifically, the stage method divides all the patients into 3 subgroups according to the TNM stage. For DGM2FS or M2FS approach, k-means clustering algorithm is adopted to aggregate the patients into 3 subgroups based on the selected features. Then, we test if these 3 subgroups has distinct survival outcome using log-rank test [2]. As can be seen from Fig. 4, the proposed DGM2FS method could achieve superior stratification performance than the competitors. These results further show the promise of incorporating diagnosis information for patient stratification from multi-modal data, which opens up the opportunity for building personalized treatment plan in the stage of cancer development.

4 Conclusion

In this paper, we develop DGM2FS, an effective multi-modal feature selection method that can identify survival associated biomarkers from both histopathological image and gene expression data. The main advantage of our approach

is its capability of utilizing the diagnosis information to guide the feature selection process, which can more accurately predict the clinical outcome for lung squamous cell carcinoma patients. DGM2FS is a general framework and can be easily transferred to other types of cancers or predict the response of a specific treatment, which opens up new opportunity in personalized treatment.

References

1. Liu, J., Lichtenberg, T.: An integrated tcga pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell **173**(2), 400–416 (2018)
2. Cheng, J., Huang, K.: Integrative analysis of histopathological images and genomic data predicts clear cell renal cell carcinoma prognosis. Cancer Res. **77**(21), 91–100 (2017)
3. Yuan, Y., Rueda, M.: Quantitative image analysis of cellular heterogeneity in breast tumors complements genomic profiling. Sci. Transl. Med. **4**(157), 143–157 (2012)
4. Yao, J., Huang, J.: Deep correlational learning for survival prediction from multi-modality data. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 406–414 (2017)
5. Shao, W., Cheng, J.: Ordinal multi-modal feature selection for survival analysis of early-stage renal cancer. In: International Conference on Medical Image Computing and Computer Assisted Intervention, pp. 648–656 (2018)
6. Phoulady, H., Dmitry, B.: Nucleus segmentation in histology images with hierarchical multilevel thresholding. In: International Conference on SPIE, pp. 1–8 (148)
7. Zhang, J., Lu, K.: Weighted frequent gene co-expression network mining to identify genes involved in genome stability. PLoS Comput. Biol. **8**(8), 1–14 (2012)
8. Zhang, Y.: A regularization approach to learning task relationships in multitask learning. ACM Trans. Knowl. Discov. Data **8**(3), 1–12 (2012)
9. Al-Shibli, K.I., Donnem, T., Al-Saad, S., Persson, M., Bremnes, R.M., Busund, L.T.: Prognostic effect of epithelial and stromal lymphocyte infiltration in non-small cell lung cancer. Clin. Cancer Res. **14**(16), 5220–5227 (2008)
10. Liu, Y.: Cancer and innate immune system interactions: translational potentials for cancer immunotherapy. J. Immunother. **35**(4), 299–299 (2012)