Kernel Learning Framework for Cancer Subtype Analysis with Multi-omics Data Integration

William Bradbury[1], Thomas Lau[2], Shivaal Roy[1]
[1]Department of Computer Science, [2]Department of Bioengineering

Background

• Supervised multiple kernel learning has been implemented in *Gevaert et al.* to classify rectal cancer microarray data. The resulting Support Vector Machine accurately classified different outcomes, often reaching above 0.90. The model also performed better when using more than one genome-widedata set, suggesting that integrating multiple genome-wide data sources allows models to reach higher accuracy.

• The iCluster algorithm developed by *Shen et al.* was used to accurately group cancer outcomes based on multi-omic considerations found in TCGA data on breast cancer. The algorithm clusters different cancer subtypes using multiple genomic features such as DNA copy number changes and gene expression. iCluster does not however implement any kernel methods.

• Doctors have independently identified cancer subtypes in patients through measuring the progression of their cancers. However, these subtype classifications are slow to perform and can only be made after the cancer has already progressed a fair amount, thereby not proving effective in determining the best method of treatment.

Objective

• **Extend** existing unsupervised multiple kernel learning technique in Zhuang et al. to leverage sparse labeling for specific use with clinical and multi-omic data from TCGA.

• **Characterize** genomic and multi-omic features of known cancer subtypes and identify previously unknown subtypes.

Problem Formulation

• Multiple Kernel Learning is the use of a linear combination of kernels to map points to a higher-dimensional feature space where they can be more easily separated by an SVM. We write this as

$$\mathcal{K}_{conv} = \sum_{t=1}^{m} \mu_t k_t$$

where u_t is the weight assigned to each kernel. In MKL, we try to learn the kernel combination that linearly separates the data best.

Unsupervised Multiple Kernel Learning

• Unsupervised Multiple Kernel Learning also implements a linear combination of kernels to create a distance metric in a higher-dimensional space. The distance metric is used to determine groupings using a k-means or alternate clustering algorithm.

Sparse Labels

• We take advantage of various labels, although sparse, to constrain our data and aid in the clustering process. We incorporate the clinical data mentioned above into our cost function to have the resultant combination of kernels reflect this added restriction. An additional constraint we impose uses our knowledge of non-cancer patients in TCGA. Our algorithm should not cluster cancer and non-cancer patients into the same group and thus we can incorporate this knowledge into our choice of kernels.

Optimization Problem

• **Cluster-label alignment metric**

- A good kernel should induce kernel values which place samples of the same subtype together in the feature mapping. In other words for each $\mathbf{x}_i$ we expect that the optimal kernel minimizes $\sum_{\mathbf{x}_j \in \mathcal{L}_i} k_{ii} - 2k_{ij} + k_{jj}$, where C_i is the set of all samples with the same known label as $\mathbf{x}_i$. $k_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$. $C_i = \varnothing$ for $\mathbf{x}_i$ that do not have known labels.
- A good kernel should induce kernel values which place samples of different subtypes apart in the feature mapping. When this is the case, we expect the kernel to maximize $\sum_{\mathbf{x}_j \in \mathcal{L}_{i \neq j}} k_{ii} - 2k_{ij} + k_{jj}$. Where here $C_{i \neq j}$ is the set of sample points labeled with a different subtype than $\mathbf{x}_i$.

3.4.2 Cost function

We combine these cluster-label alignment metrics with those found in *Zhuang et al.* to form the overall optimization:

$$\min_{\mathcal{B}, k \in \mathcal{K}_{conv}} \frac{1}{2} \sum_{i=1}^{n} \left\| \mathbf{x}_i - \sum_{\mathbf{x}_j \in B_j} k_{ij} \mathbf{x}_j \right\|^2 + \gamma_1 \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in B_i} k_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|^2 + \gamma_2 \sum_i |\mathcal{B}_i| + \gamma_3 \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \mathcal{L}_i} (k_{ii} - 2k_{ij} + k_{jj}) - \gamma_4 \sum_{i=1}^{n} \sum_{\mathbf{x}_j \in \mathcal{L}_{i \neq j}} (k_{ii} - 2k_{ij} + k_{jj})$$

where

$\mathcal{B}$ is a clustering assignment,

$\mathcal{K}_{conv}$ is given by $\{k(\cdot,\cdot) = \sum_{t=1}^{m} \mu_t k_t(\cdot,\cdot) : \sum_{t=1}^{m} \mu_t = 1, \mu_t \geq 0\}$,

γ_i are meta-parameters controlling trade-offs between constraints, and

$\mathcal{L}$ is given above.

3.5 A simpler formulation

As in *Zhuang et al.* we define matrices $\mathbf{D}, \mathbf{S} \in \{0,1\}^{n \times n}$ where each element

$$[\mathbf{D}]_{ij} = \mathbb{1}\{\mathbf{x}_j \in \mathcal{B}_i\}$$
$$[\mathbf{S}]_{ij} = \mathbb{1}\{\mathbf{x}_j \in \mathcal{L}_i\}$$
$$[\mathbf{M}]_{ij} = \mathbf{x}_i^T \mathbf{x}_i + \mathbf{x}_j^T \mathbf{x}_j - 2\mathbf{x}_i^T \mathbf{x}_j$$
$$[\mathbf{Q}]_{ij} = \mathbb{1}\{\mathbf{x}_j \in \mathcal{L}_{i \neq j}\}$$

so that we can write:

$$\min_{\mu, \mathbf{D}} \frac{1}{2} \|\mathbf{X}(\mathbf{I} - \mathbf{K} \circ \mathbf{D})\|_F^2 + \gamma_1 \mathrm{tr}\, \mathbf{K} \circ \mathbf{D} \circ \mathbf{M}(\mathbf{1}\mathbf{1}^T) + \gamma_2 \|D\|_{1,1} + \mathrm{tr}\left(\left[(\mathbf{K} \circ \mathbf{I})(\mathbf{1}\mathbf{1}^T) - 2\mathbf{K} + (\mathbf{1}\mathbf{1}^T)(\mathbf{K} \circ \mathbf{I})^T \right] \circ (\gamma_3 \mathbf{S} + \gamma_4 \mathbf{Q}) \right) (\mathbf{1}\mathbf{1}^T)$$
$$\text{s.t } [\mathbf{k}]_{ij} = \sum_{t=1}^{m} \mu_t k^t(\mathbf{x}_i, \mathbf{x}_j), 1 \leq i, j \leq n, \mu^T \mathbf{1} = 1, \mu \geq 0, \mathbf{D} \in \{0,1\}^{n \times n}$$

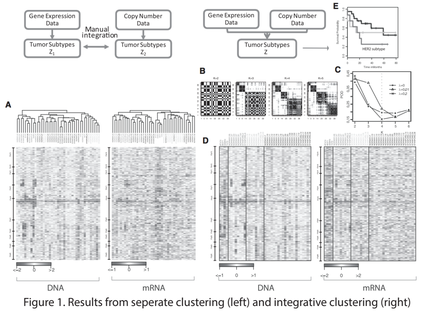

Figure 1. Results from seperate clustering (left) and integrative clustering (right)

4 Alternating Optimization Algorithm

4.1 Solving μ by fixing D

Fixing $\mathbf{D}$ we try to solve for μ by minimizing the cost function:

$$J(\mu) = \mu^T \left(\sum_{t=1}^{m} \sum_{i=1}^{n} \mathbf{k}_{t,i} \mathbf{k}_{t,i}^T \circ \mathbf{d}_i \mathbf{d}_i^T \circ P \right)^T \mu + \mathbf{z}^T \mu,$$

where

$$[\mathbf{z}]_t = \sum_{i=1}^{n} (2\gamma_1 \mathbf{v}_i \circ \mathbf{d}_i - 2\mathbf{p}_i \circ \mathbf{d}_i + \gamma_3 (\mathbf{1}_i \sum_j S_{ij} + \mathbf{1}_i \sum_j S_{ji} - 2\mathbf{s}_i) + \gamma_4 (\mathbf{1}_i \sum_j Q_{ij} + \mathbf{1}_i \sum_j Q_{ji} - 2\mathbf{q}_i))^T \mathbf{k}_{t,i}$$
$$\mathbf{P} = \mathbf{X}^T \mathbf{X}$$
$$\mathbf{k}_{t,i} = [k^t(\mathbf{x}_i, \mathbf{x}_1), ..., k^t(\mathbf{x}_i, \mathbf{x}_n)]^T$$
$\mathbf{p}_i$ is the i-th column of $\mathbf{P}$
$\mathbf{v}_i$ is the i-th column of $\mathbf{M}$
$\mathbf{s}_i$ is the i-th column of $\mathbf{S}$
$\mathbf{q}_i$ is the i-th column of $\mathbf{Q}$

Most importantly, this cost function is a Quadratic Program and so is relatively efficient to solve compared to the previous form.

4.2 Solving D by fixing the kernel K

The optimization step for $\mathbf{D}$ is unchanged from its presentation in *Zhuang et al.* and so we simply give the result here:

$$J(\mathbf{d}) = \mathbf{d}^T \left(\mathbf{k}\mathbf{k}^T \circ \mathbf{P} \right) \mathbf{d} + (2\gamma_1 \mathbf{k} \circ \mathbf{v} - 2\mathbf{k} \circ \mathbf{p})^T \mathbf{d},$$

which is of the form

$$J(\mathbf{d}) = \mathbf{d}^T \mathbf{W} \mathbf{d} + \mathbf{c}^T \mathbf{d}$$

and so is also tractable with a QP convex optimization package.

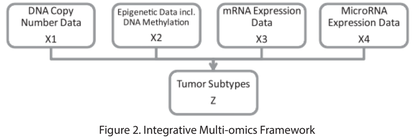

Figure 2. Integrative Multi-omics Framework

Data sources

• **TCGA**
The Cancer Genome Atlas is the world's largest collection of multi-omic cancer data, collection from US patients by hospitals around the world.

• **Clinical Data**
TCGA contains sparse clinical data including eventual patient outcomes, optimal treatment plans, and clinicial established subtypes.

• **Multi-omic Data**
Multi-omic data available in TCGA includes:
- Copy Number Variation
- Methylation Data
- miRNA
- mRNA
- RPPA

• **Patient Class Data**
Each patient is tagged with patient information, specifically which hospital they were treated at and whether or not they had cancer. This can be used as another source of sparsely labeled data in the clustering process above.

iCluster

The iCluster paper by *Shen et al.* is a source of reproducible and standard automated subtype analyses. These can be used both as validation and as seed labels for our algorithm. In the first case it can be used to demonstrate demonstrate robustness of this algorithm while in the second case it can be used to jump start the process of discovering subtypes.

Acknowledgements

We thank Prof. Olivier Gevaert for his help in formulating and guiding this project and for his help in using TCGA and MKL. We thank Prof. Serafim Batzoglou for his suggestions which led us to cancer subtype analysis in thefirst place.

Literature Cited

[1] Anneleen Daemen et al. "A kernel-based integration of genome-wide data forclinical decision support." In:Genome medicine1.4 (2009), p. 39.

[2] Anneleen Daemen et al. "Improved microarray-based decision support with graph encoded interactome data." In:PloS one5.4 (2010), e10225.

[3] GRG Lanckriet and Nello Cristianini. " Learning the kernel matrix with semidef-inite programming". In:Journal of Machine Learning Research5 (2004),pp. 27–72.

[4] Cheng Li and Ariel Rabinovic. "Adjusting batch effects in microarray ex-pression data using empirical Bayes methods". In:Biostatistics8.1 (2007),pp. 118–127.

[5] R. Shen, A. B. Olshen, and M. Ladanyi. "Integrative clustering of multiplegenomic data types using a joint latent variable model with application tobreast and lung cancer subtype analysis". In:Bioinformatics25.22 (2009),pp. 2906–2912.

[6] Ronglai Shen et al. "Integrative Subtype Discovery in Glioblastoma UsingiCluster". In:PLoS ONE7.4 (2012), e35236.

[7] Emily a. Vucic et al. "Translating cancer 'omics' to improved outcomes". In:Genome Research22.2 (2012), pp. 188–195.

[8] Jinfeng Zhuang et al. "Unsupervised multiple kernel learning". In:Proceed-ings of the Third Asian Conference on Machine Learning20 (2011), pp. 129–144.