Maximizing Information Through Multiple Kernel-based Heterogeneous Data Integration and Applications to Ovarian Cancer

Jaya Thomas
Department of Computer Science
State University of New York, Incheon, Korea and
Stony Brook University, New York, USA
jaya.thomas@sunykorea.ac.kr

Lee Sael
Department of Computer Science
State University of New York, Incheon, Korea and
Stony Brook University, New York, USA
sael@cs.stonybrook.edu

ABSTRACT

The medical research facilitates to acquire diverse type of data from the same individual for a particular cancer. Recent studies show that utilizing such diverse data results in more accurate predictions. The major challenge faced is how to utilize such diverse data sets in an effective way. In this paper, we introduce a multiple kernel based pipeline for integrative analysis of high-throughput molecular and clinical data. We apply the pipeline on Ovarian cancer data from TCGA. After multiple kernel have been generated from weighted sum of individual kernels, it is used to stratify patients and predict clinical outcomes. We examine the clinical outcomes of each subtype to verify how well they cluster.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications – *Data mining;* I.2.6 [**Artificial Intelligence**]: Learning – *Parameter learning*

General Terms

Algorithms.

Keywords

Integrative analysis; Multiple kernel; Molecular data; Clinical data; Patient stratification

1. INTRODUCTION

Ovarian cancer is the fifth most common cancers diagnosed in females with overall five year survival rate only around 44%. The Cancer Genome Atlas (TCGA) [9] reports diverse genomic information with paired clinical information for more than 500 cases of ovarian serous cystadenocarcinoma. In cancer data analysis, including ovarian cancer data, a stratification can be improved by integrative analysis of the multiple bio-clinical data. However, due to the complex relationship between the multiple data types, the integrative analysis is still a challenging task.

The patient stratification is to find subgroups of patients to allow better detection and interpretation as well as predict outcomes in specific subgroup. Kim et al. [4] considers somatic mutation profile and exploited k-means clustering to identify the tumor subtypes. In their recent work [5], a compressed somatic mutation profile was suggested for fast comparison. Hofree et al. [2] has used genome-scale somatic mutation profiles in combination with a gene interaction network to carry out subgrouping of patients. Recently, Wang et al. [12] proposed a modified consensus clustering to carry out patient stratification for breast cancer patients.

Analysis of one or few data types may not be sufficient for accurate stratification. Thus, efforts to integrate the molecular data were carried out. Thomas et al. [10] work presents two general class of heterogeneous data integration, i.e., Multiple Kernel learning and Bayesian network. Kim et al. [3] proposed a graph based integrated framework using four genome data types to carry out molecular based classification of clinical outcomes. Sohn et al. [8] modeled the influence of multi-layered genomic features on gene expression traits by modeling an integrative statistical framework based on a sparse regression. Schafer et al. [7] integrated copy number and gene expression by a modified correlation coefficient and an explorative Wilcoxon test to find DNA regions of abnormalities. Mankoo et al. [6] have applied multivariate Cox Lasso model and median time-to-event prediction algorithm on data set integrated from the four genomic data. Yuan et al. [13] evaluated the predictive power of patient survival and binary clinical outcome using clinical data in combination with one molecular data..

Integrative analysis method that can cover heterogeneity of data types in molecular data and clinical data can beneficial in predicting the prognostics of patients via stratifying the patients in the different risk groups. Multiple kernel learning is well known for addressing various data heterogeneity. Moreover, Kernel methods, including multiple kernels, are well suited handling non-linearity of high dimensional data by mapping data to feature space.

EDB2016, October 17-19, 2016, Jeju Island, Korea.
© 2016 ACM. ISBN 978-1-4503-4754-9 …$15.00.
DOI: http://dx.doi.org/10.1145/3007818.3007831

In this paper, we make the following contributions:

1. **Combines clinical data with multiple molecular data**. We examine how adding more molecular information increases the prediction performance in stratifying ovarian cancer patients, and predicting tumor grade and patient survival time.

2. **Propose a multiple kernel based pipeline model to integrate multiple heterogeneous data types.** The proposed model allows to analyze heterogeneous data i.e., combines data with diverse background distributions, relations, dimensions, and formats to enhance the statistical significance and thus, obtain more refined information.

2. MATERIALS AND METHODS

2.1 Datasets and Raw Mutation Scores

Data are initially selected and downloaded 312 samples that contained all four genomic data types, i.e., copy number alternation, methylation, mRNA expression and the mutation information, from TCGA data portal [16] via TCGA assembler [14]and TCGA Firehose [17]. Clinical information of the 312 samples is also downloaded from TCGA. The clinical data includes the survival time (days to death), age, tumor stage, tumor grade, vital status and neoplasm cancer status.

2.2 Kernel Matrix Representing Molecular Information

The patient-to-gene set matrix of the four data sources are used to create kernels using kernel functions. A feature function, ϕ(x), maps the original data feature x in the input space to a high-dimensional feature space. A Kernel function is a function that corresponds to the inner product in an expanded feature space. The size of a kernel matrix is independent of the number of features and is solely dependent on the number of data. In practice, an explicit definition of feature function, ϕ(x), is not needed since they are tightly integrated in the definition of the kernel functions.

The kernel functions we used are linear and radial basis function (RBF). We explored the use of commonly used kernels including linear, sigmoid, polynomial and the radial basis function. We chose the kernel function that showed the best performance for each data type.

2.3 Multiple Kernel Learning for Cancer Classification

The kernel matrix constructed from each data types is further integrated to form a single kernel matrix using a multiple kernel learning approach. Several methods are suggested for integrating the kernels [1]. We take a two-step approach that first combines the kernels in a weighted linear fashion and then perform learning on the combined kernel. The kernel combination is defined as follows:

$$K_\beta(\mathbf{x}_i, \mathbf{x}_j) = \sum_{s=1}^{S} \beta_s \, k_s(\phi(\mathbf{x}_i^s), \phi(\mathbf{x}_j^s))$$

$$\text{subjected to } \beta_s \geq 0 \text{ and } \sum_{s=1}^{S} \beta_s = 1,$$

where S is the number of kernels, $\mathbf{x}_i^s$ is the original feature vector of kernel s of sample *i*, and β_n is the kernel coefficient of kernel s.

To obtain optimal weights for kernel combination, we take the optimization approach suggested by Zien et al. [15]. In their approach, the kernel coefficient is determined by the efficacy of each of the kernel matrix containing sets learned by Least Square Support Vector Machine (LS-SVM). LS-SVMs are closely related to regularization networks and Gaussian processes but additionally emphasize and exploit primal-dual interpretations from the optimization theory [11]. The primal form of a LS-SVM is optimized by the following minimization problem:

$$\min_{\mathbf{w},b,err} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + \gamma \sum_{i=1}^{N} err_i^2\right)$$

$$\text{subjected to } y_i[\mathbf{w}^T\phi(\mathbf{x}_i) + b] = 1 - err_i^2 \text{ for } i = 1,2,\ldots,N$$

where w is the weight vector we are trying to learn, err_s is the error variables that represent the value corresponding to misclassification in case of overlapping distribution, and γ is the regularization parameter that tackles data over fitting problem.

2.4 Stratification Using Kernel K-means

Stratification of patients can be done with clustering methods. We use kernel K-means on the generated multiple kernel matrix for stratifying the Ovarian cancer to subtypes. The multiple kernel matrix contains the similarity information about pairs of data in the combined feature space. Thus, when we apply the kernel k-means to the multiple kernel matrix, data are clustered so that the clustering error is minimized in the combined feature space.

3. RESULTS

We report the results for validation and performance of combining the clinical features with the biological features by the multiple-kernel on two important translational bioinformatics tasks: patient stratification and clinical predictions.

3.1 Patient Stratification via K-means

We performed the kernel k-means clustering to stratify ovarian cancer patients using the generated kernel matrices. We compared four data type combinations as input to the k-mean clustering: the first multiple kernel is constructed from only the molecular data types, the second is constructed from clinical information (i.e., age, stage, grade), the third is constructed by non-weighted linear combination of kernels of molecular as well as clinical data, and the fourth is construed by weighted linear combination of kernels of molecular and clinical data.

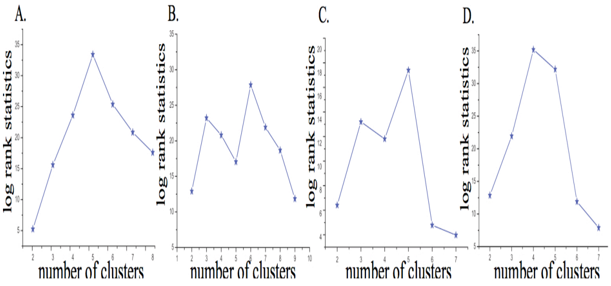

Figure 1. Log rank statistic to determine the number of clusters (A) Molecular data (B) Clinical data (C) Molecular and clinical data with non-weighted linear kernel coefficient

(D) Molecular and clinical data with weighted kernel coefficient.

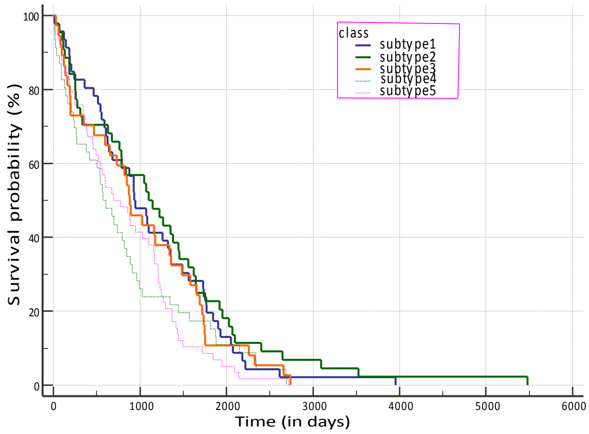

Figure 2. Kernel k-means clustering of the TCGA OV all molecular data.

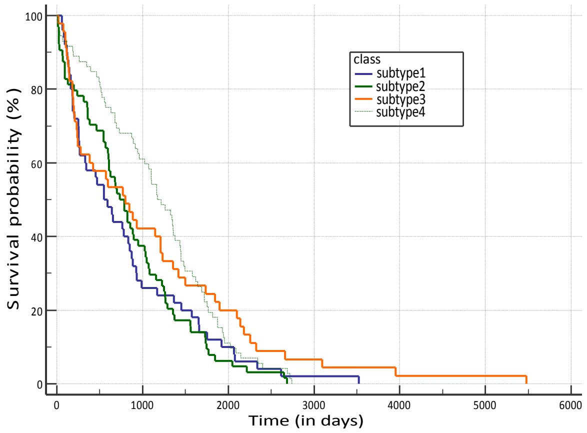

Figure 3. Kernel k-means clustering for clinical data

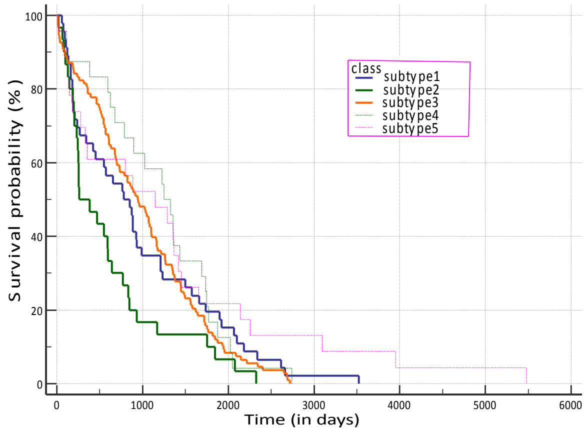

Figure 4. Kernel k-means clustering of the TCGA OV all molecular data with clinical data with linear kernel combination

To evaluate the clustering result, we performed survival analysis on each clusters, or subgroups, using the Cox proportional hazards regression model in the R survival package for each of the data type combinations. Out of 312 patient samples, the clustering was carried out for 75% (231) of the samples and 25% (81) to determine the number of clusters, k.

The value of k (i.e., the number of clusters) was determined using the log rank statistics. Figure 1 shows the different log rank statistic values obtained for different number of clusters. Figure 1 (A) shows the plot for integrated molecular data indicating the best value for k being 5. Figure 1(B) is a graph for determining the k (i.e., 5) value for clinical data. Similarly, figure 1(C) and figure 1(D) are plots when molecular data is integrated with clinical without and with weighted kernel coefficient, results in best clusters for k=5 and k=6 respectively. We compared the survival times for these clusters using log-rank statistics and obtained the P-value. The P-value for all the above cases is less than 0.05. Thus, it shows that there exists a significant separation between the subgroups with respect to survival time.

Setting k=5 in kernel k-means clustering, the p-value of the subtype separation for survival analysis is 0.02 for all molecular data types (figure 2), 0.0079 for clinical data (figure 3), 0.009 for integrated molecular and clinical data with linear kernel coefficient (figure 4). It can be observed that the clusters identified by integrating the clinical data are more predictive with log-rank p-value of 1.4 x 10^{-3}. The size of the cluster formed is not uniform; however the method shows an ability to categorize the patient samples into sub groups that significantly differ in the survival time.

4. CONCLUSION

In this paper, we have developed a multiple kernel learning based pipeline for integrative analysis of heterogeneous data types and apply it on ovarian cancer data. The data types we look at are molecular data and clinical data. The model is used to carry out patient stratification. We use kernel k-means to perform stratification of patient for survival time, and tumor grade. Stratification is done considering different test cases including integrated molecular data, clinical data, and integration using linear combination. The patient stratification results for different test cases show that the integration of molecular and clinical data results in better pattern forming relation.

ACKNOWLEDGMENTS

This research was supported by NRF of Korea funded by the MSIP (NRF-2015R1C1A2A01055739), MSIP supervised by IITP (IITP-2016-R0346-15-1007) and by the KEIT, Korea, under the "Global Advanced Technology Center" (10053204).

REFERENCES

[1] Mehmet Gonen and Ethem Alpaydın. 2011. Multiple kernel learning algorithms. *Journal of Machine Learning Research* 12: 2211–2268.

[2] Matan Hofree, John P Shen, Hannah Carter, Andrew Gross, and Trey Ideker. 2013. Network-based stratification of tumor mutations. *Nature Methods* 10, 11: 1108–1115. https://doi.org/10.1038/nmeth.2651

[3] D Kim, H Shin, Y S Song, and J H Kim. 2012. Synergistic

effect of different levels of genomic data for cancer clinical outcome prediction. *J Biomed Inform* 45: 1189–1191.

[4] Sungchul Kim, Lee Sael, and Hwanjo Yu. 2014. Identifying cancer subtypes based on somatic mutation profile. In *Proceedings of the 8th International Workshop on Data and Text Mining in Biomedical Informatic - DTMBIO'14*.

[5] Sungchul Kim, Lee Sael, and Hwanjo Yu. 2015. A mutation profile for top- k patient search exploiting Gene-Ontology and orthogonal non-negative matrix factorization. *Bioinformatics*: btv409.

[6] Parminder K Mankoo, Ronglai Shen, Nikolaus Schultz, Douglas A Levine, and Chris Sander. 2011. Time to recurrence and survival in serous ovarian tumors predicted from integrated genomic profiles. *PLoS One* 6, 11: e24709. https://doi.org/10.1371/journal.pone.0024709

[7] M Schafer, H Schwender, S Merk, C Haferlach, K Ickstadt, and M Dugas. 2009. Integrated analysis of copy number alterations and gene expression: a bivariate assessment of equally directed abnormalities. *Bioinformatics* 25: 3228–3235.

[8] K A Sohn, D Kim, J Lim, and J H Kim. 2013. Relative impact of multi-layered genomic data on gene expression phenotypes in serous ovarian tumors. *BMC systems biology* 7: S9.

[9] TCGA The Cancer Genome Atlas Research Network. 2011. Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 7353: 609–15. https://doi.org/10.1038/nature10166

[10] Jaya Thomas and Lee Sael. 2015. Overview of integrative analysis methods for heterogeneous data. In *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, 266–270.

[11] J. Vandewalle and J A K Suykens. 1999. Least squares support vector machine classifiers. *Neural Processing Lett* 9: 293–300. https://doi.org/10.1023/A:1018628609742

[12] C Wang, R Machiraju, and K Huang. 2014. Breast cancer patient stratification using a molecular regularized consensus clustering method. *Methods* 67(3): 304–312.

[13] Yuan Yuan, Eliezer M Van Allen, Larsson Omberg, Nikhil Wagle, Ali Amin-Mansour, Artem Sokolov, Lauren a Byers, Yanxun Xu, Kenneth R Hess, Lixia Diao, Leng Han, Xuelin Huang, Michael S Lawrence, John N Weinstein, Josh M Stuart, Gordon B Mills, Levi a Garraway, Adam a Margolin, Gad Getz, and Han Liang. 2014. Assessing the clinical utility of cancer genomic and proteomic data across tumor types. *Nature Biotechnology* 32, 7: 644–652.

[14] Y Zhu, P Qiu, and Y Ji. 2008. TCGA-Assembler: Open-Source Software for Retrieving and Processing TCGA Data. *Nature Methods* 11(6): 599–600. https://doi.org/10.1038/nmeth.2956

[15] Alexander Zien and Cheng Soon Ong. 2007. Multiclass multiple kernel learning. In *The 24th International Conference on Machine Learning*, 1191–1198. https://doi.org/10.1145/1273496.1273646

[16] The Cancer Genome Atlas. http://tcga-data.nci.nih.gov/.

[17] Firehose Broad GDAC. http://gdac.broadinstitute.org/.