Genome-Wide Canonical Correlation Analysis-Based Computational Methods for Mining Information from Microbiome and Gene Expression Data

Rayhan Shikder[1,2], Pourang Irani[1], and Pingzhao Hu[1,2(✉)]

[1] Department of Computer Science, University of Manitoba, Winnipeg, Canada
shikderr@myumanitoba.ca,
pourang.irani@cs.umanitoba.ca,
pingzhao.hu@umanitoba.ca
[2] Department of Biochemistry and Medical Genetics,
University of Manitoba, Winnipeg, Canada

Abstract. Multi-omics datasets are very high-dimensional in nature and have relatively fewer number of samples compared to the number of features. Canonical correlation analysis (CCA)-based methods are commonly used for reducing the dimensions of such multi-view (multi-omics) datasets to test the associations among the features from different views and to make them suitable for downstream analyses (classification, clustering etc.). However, most of the CCA approaches suffer from lack of interpretability and result in poor performance in the downstream analyses. Presently, there is no well-explored comparison study for CCA methods with application to multi-omics datasets (such as microbiome and gene expression datasets). In this study, we address this gap by providing a detail comparison study of three popular CCA approaches: regularized canonical correlation analysis (RCC), deep canonical correlation analysis (DCCA), and sparse canonical correlation analysis (SCCA) using a multi-omics dataset consisting of microbiome and gene expression profiles. We evaluated the methods in terms of the total correlation score, and the classification performance. We found that the SCCA provides reasonable correlation scores in the reduced space, enables interpretability, and also provides the best classification performance among the three methods.

Keywords: Canonical correlation analysis (CCA) · RCC · DCCA · SCCA · Comparison study · Multi-omics data · Microbiome and gene expression data

1 Introduction

Datasets comprising of multiple feature sets from different omics sources (e.g., genomics, proteomics, microbiomics etc.) measured on the same subjects are known as multi-omics (or multi-view) data. Integrated study of the multi-omics data has the potential to reveal more information about a disease as it may tell us about the individual associations, interactions among the factors and the flow of information from cause of the disease to consequences [1]. However, most of the omics datasets are very

M.-J. Meurs and F. Rudzicz (Eds.): Canadian AI 2019, LNAI 11489, pp. 511–517, 2019.
https://doi.org/10.1007/978-3-030-18305-9_53

high dimensional in nature and combining them usually results in a unique representation where the numbers of features are very large (e.g., tens of thousands) compared to the number of available samples (e.g., hundreds). These large number of features create challenges in applying most of the statistical methods. Moreover, a large subset of these features may represent redundant or irrelevant information. Therefore, prior to learning any objective functions or finding associations among the omics datasets, the feature sets need to be reduced to a lower dimensional subspace.

Most often, researchers want to investigate the relationships between two omics datasets. Canonical correlation analysis (CCA) - based approaches, which finds the linear combinations of features from two datasets and tries to maximize the correlation between them, are common ways to find such relationships [2]. In addition, CCA also reduces the dimensionality of the original high-dimensional omics datasets, making it suitable for fusion and downstream predictive analysis. However, in a setting where the numbers of features outnumber the number of samples, the basic version of the CCA is not effective. To deal with this situation, regularized versions of the canonical correlation analysis (regularized canonical correlation analysis or RCCA) have been developed [3, 4]. To learn non-linear combinations of the features while calculating the correlations, deep neural network based parametric version of the CCA (named as deep canonical correlation analysis (DCCA)) has been proposed too [5, 6]. In biological applications, researchers also seek to trace the original features that correspond to the resulting correlations, which is hard to achieve with either RCC or DCCA. Hence, sparse versions of the canonical correlation analysis (SCCA) methods have been developed [7–9]. However, there exist no study which highlighted the comparison of the approaches with application to multi-omics datasets specially datasets consisting of microbiome and gene expression profiles.

Our contribution in this paper includes a detailed comparison of the canonical correlation methods (RCC, SCCA, and DCCA) in terms of correlation score and classification performance with applications to a multi-omics dataset consisting of microbiome and gene expression profiles. To the best of our knowledge, this study is the first to investigate the CCA approaches for microbiome and gene expression data together.

2 Preliminaries

2.1 Canonical Correlation Analysis (CCA)

Having two datasets X_1 and X_2 with $(n \times p_1)$ and $(n \times p_2)$ dimensions measured on the same subject $i = 1, 2 \ldots, n$, CCA finds linear combinations of the features from the two datasets which are maximally correlated [2]. In other words, CCA finds the linear projections $w_1^T X_1$ and $w_2^T X_2$ which have a maximum correlation between them, where w_1 and w_2 are the canonical coefficients. Let $\sum_{11}$, and $\sum_{22}$ be the covariances of X_1 and X_2, and $\sum_{12}$ be the cross-covariance between the features of the datasets, then the objective of the CCA method is to maximize the following:

$$corr\left(w_1^T X_1, w_2^T X_2\right) \quad or, \quad \left(\frac{w_1^T \sum_{12} w_2}{\sqrt{w_1^T \sum_{11} w_1 w_2^T \sum_{22} w_2}}\right) \tag{1}$$

2.2 Regularized Canonical Correlation (RCC) Analysis

When the number of features (p_1 or p_2) become larger than the total number of samples (n), the basic version of the CCA doesn't work as the first n canonical variates possess larger values while the rest of the canonical covariates becomes zero [10]. To deal with this, regularization parameters (λ_1 and λ_2) can be added with the covariance matrices in the following manner (I_{p_1} and I_{p_2} are identity matrices) [3, 4].

$$\sum\nolimits'_{11} = \sum\nolimits_{11} + \lambda_1 I_{p_1} \tag{2}$$

$$\sum\nolimits'_{22} = \sum\nolimits_{22} + \lambda_2 I_{p_2} \tag{3}$$

2.3 Deep Canonical Correlation Analysis (DCCA)

DCCA finds complex nonlinear projections of the input features which are maximally correlated [5]. DCCA is a deep neural network (DNN)-based approach, where two densely connected networks (Network 1 and Network 2 in Fig. 1) are separately trained on two views of the dataset. These two networks learn nonlinear feature combinations and use a correlation maximization objective function.

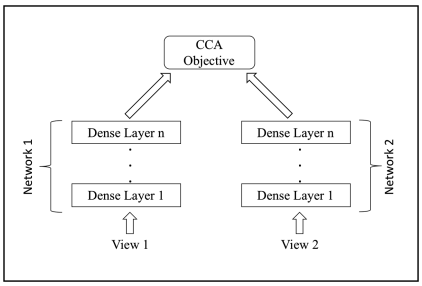

Fig. 1. Schematic diagram of the deep canonical correlation analysis (DCCA) method

2.4 Sparse Canonical Correlation Analysis (SCCA)

For datasets with large number of features, the interpretation of linear combinations become impracticable. In such cases, considering a sparse subset of the features is a viable approach [8, 9]. In this case, the objective function to be maximized takes the following form:

$$corr\left(w_1^T X_1, w_2^T X_2\right)$$
$$where\ ||w_1||^2 \leq 1, ||w_2||^2 \leq 1, \boldsymbol{P}_1(w_1) \leq c_1,\ and\ \boldsymbol{P}_2(w_2) \leq c_2 \tag{4}$$

Here, $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ are called penalty functions or sparse CCA criterion. These penalty functions are chosen in a way to provide sparse feature combinations and also to make the CCA deal with situations where the feature sets are large compared to the number of samples. $\boldsymbol{P}_1$ and $\boldsymbol{P}_2$ can be lasso or fused lasso penalty functions. The parameters c_1, c_2 are used to control the level of penalization.

3 Experiments and Results

3.1 Dataset

We considered a multi-omics dataset consisting of two views: gene expression and microbiome profiles [11]. There are 184 samples with 4 disease subtypes. The gene expression profiles of the data consist of 20,253 features, each of which represents the level of expression for a particular gene. The microbiome profiles consist of 7,000 features which are sparse, discrete in nature and represent the count of an operational taxonomic unit (OTU) in the sample.

3.2 Preprocessing and Hyperparameter Tuning

All the features with no variation (such as zero and constant values) across all of the 184 samples were removed from the dataset. The remaining dataset contained 20,251 gene features and 5,443 OTU features which were normalized afterwards. The 184 samples were divided into train (147) and test (37) groups in a stratified manner. We used R package: CCA for the RCC [10], python implementation from [12] for DCCA, and the PMA package in R [13] for SCCA. We have also considered a supervised version of the SCCA approach (we call this SCCA(S)), where the learned feature projections are also correlated with the output labels. For all the methods, we tuned the hyperparameters to their appropriate values using the training set.

3.3 Total Correlation Scores

After tuning the hyperparameters, we performed the canonical correlation analyses for different output dimensions and learned the canonical coefficients (w_1 and w_2). We multiplied these coefficients with the original dataset to generate the projections.

Total correlation scores were calculated (using the *linear_cca* method provided in [12]) from the projections. From Fig. 2, we can see that RCC provides better correlation scores. On the other hand, SCCA, SCCA(S), and DCCA provide almost similar correlation scores. With the increasing number of output dimensions, the correlation scores become almost the same for all of the approaches when the output dimension surpasses the number of test samples. For SCCA, the sparsity nature may correspond to the compromise in the total correlation score. As deep neural network (DNN)-based

approaches are always data hungry, the fewer number of samples is the main reason behind the relatively lower correlation scores of the DCCA method.

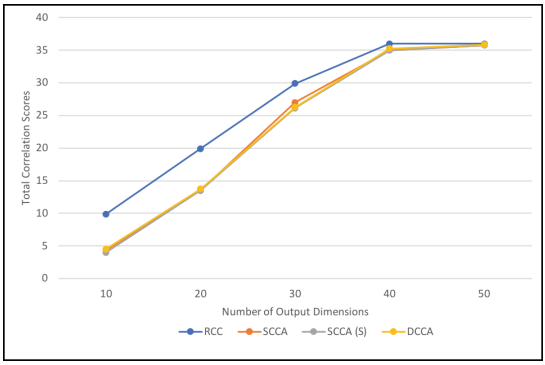

Fig. 2. Total correlation scores for different canonical correlation approaches.

3.4 Classification Performance

We performed binary classifications using support vector machine (SVM) on the projected data from the CCA methods. The classification labels were converted into binary by keeping one class in one group and all the others in another. Hyperparameters (kernels, C, sigma, gamma etc.) of the SVM method were adjusted.

Table 1. Binary-class classification results using SVM on the output projections from different CCA methods. Evaluation metrics are accuracy and area under the ROC curve (AUC).

Dimensions	Metrics	RCC	SCCA	SCCA (S)	DCCA
10	Accuracy	67.56%	72.97%	72.97%	67.56%
	AUC	0.5	0.6	0.6	0.5
20	Accuracy	67.56%	75.67%	75.67%	67.56%
	AUC	0.5	0.71	0.67	0.5
30	Accuracy	70.27%	**75.67%**	**75.67%**	67.56%
	AUC	0.54	**0.756**	**0.67**	0.5
40	Accuracy	70.27%	**78.37%**	**78.37%**	67.56%
	AUC	0.54	**0.69**	**0.69**	0.5
50	Accuracy	67.56%	70.27	70.27	67.56%
	AUC	0.5	0.63	0.6	0.5

From Table 1, we can see that DCCA provides the worst classification performances. The smaller sample size of the dataset may be the reason behind this performance loss. The RCC method's performance is also poor which is easily observed with the low AUC values. Although multi-omics datasets are very high-dimensional,

only a handful of these dimensions are actually responsible for a particular phenotype. Therefore, incorporating all the input features may be the reason for the poor classification performance of RCC. Finally, it is visible that the SCCA methods provide relatively better classification performances. The sparse nature of these methods may be the main reason behind this. However, it is surprising that the supervised version of the SCCA didn't provide any better results than the unsupervised one.

4 Conclusion

In this study, we found that sparse canonical correlation analysis provides interpretable correlation scores and better performance in downstream analysis. The regularized canonical correlation analysis, although provides good correlation scores, lacks interpretability and provides poor classification performance. On the other hand, the deep canonical correlation analysis provides moderate correlation scores but lacks interpretability and suffers from poor performance in classification. Therefore, it is advised not to use DCCA with high-dimensional multi-omics datasets having fewer numbers of samples. In future, we will run the comparisons using a larger dataset. We hypothesize that incorporating the outcome variables in the DCCA approach may aid in better performance in the downstream analyses, which we will investigate in future.

Acknowledgements. This work was supported in part by Natural Sciences and Engineering Research Council of Canada, Manitoba Health Research Council and University of Manitoba.

References

1. Hasin, Y., Seldin, M., Lusis, A.: Multi-omics approaches to disease Genome Biol. **18**(1), 83 (2017)
2. Hotelling, H.: Relations between two sets of variates. Biometrika **28**, 321–377 (1936)
3. Vinod, H.D.: Canonical ridge and econometrics of joint production. J. Econom. **4**(2), 147–166 (1976)
4. Leurgans, S.E., Moyeed, R.A., Silverman, B.W.: Canonical correlation analysis when the data are curves. J. R. Stat. Soc. Ser. B. **55**(3), 725–740 (1993)
5. Andrew, G., Arora, R., Bilmes, J.A., Livescu, K.: Deep canonical correlation analysis. In: ICML (2013)
6. Wang, W., Arora, R., Livescu, K., Bilmes, J.: On deep multi-view representation learning. In: International Conference on Machine Learning, pp. 1083–1092 (2015)
7. Hardoon, D.R., Shawe-Taylor, J.: Sparse canonical correlation analysis. Mach. Learn. **83**(3), 331–353 (2011)
8. Parkhomenko, E., Tritchler, D., Beyene, J.: Sparse canonical correlation analysis with application to genomic data integration. Stat. Appl. Genet. Mol. Biol. **8**(1), 1–34 (2009)
9. Witten, D.M., Tibshirani, R., Hastie, T.: A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. Biostatistics **10**(3), 515–534 (2009)
10. Gonzalez, I., Déjean, S., Martin, P., Baccini, A.: CCA: an R package to extend canonical correlation analysis. J. Stat. Softw. **23**(12), 1–14 (2008)

11. Morgan, X.C., et al.: Associations between host gene expression, the mucosal microbiome, and clinical outcome in the pelvic pouch of patients with inflammatory bowel disease. Genome Biol. **16**(1), 67 (2015)
12. Noroozi, V.: VahidooX/DeepCCA. https://github.com/VahidooX/DeepCCA
13. Witten, D., Tibshirani, R., Gross, S., Narasimhan, B., Witten, M.D.: Package 'pma'. Genet. Mol. Biol. **8**, 28 (2013)